DETECTING STUDENT DROPOUTS USING FUZZY INFERENCING

Shahriar Husainy

Columbus State University

The College of Business and Computer Science

The Graduate Program in Applied Computer Science

Detecting student dropouts using fuzzy inferencing

A Thesis in

Applied Computer Science

by

Shahriar Husainy

Submitted in Partial Fulfillment

of the Requirements

for the Degree of

Master of Science

May 2013

I have submitted this thesis in partial fulfillment of the requirements for the degree of Master of Science.

5/29/2013

Date

Shahriar Husainy

We approve the thesis of Shahriar Husainy as presented here:

5/29/2013

Date

Shamim Khan, Ph.D.

Professor of Computer Science

Thesis Advisor

5/29/2013

Date

Lydia Ray, Ph.D.

Associate Professor of Computer Science

5/29/2013

Date

Yesem K. Peker, PhD.

Assistant Professor of Computer Science

Abstract

Fuzzy logic provides a methodology for reasoning using imprecise rules and assertions. Fuzzy inference is the process of formulating the mapping from a given input to an output using fuzzy logic. The mapping then provides a basis from which decisions can be made, or patterns discerned. This study concerns the development of a Fuzzy Inference System (FIS) for identifying likely student dropouts at Columbus State University (CSU). The fuzzy inference based model uses a hybrid knowledge extraction process to predict how likely each freshman student will be to drop their program of study at the end of their first semester. This process uses both a top down (symbolic) and a bottom-up (data-based) approach. Historical student records data have been used to evaluate the developed FIS. Findings of this study indicate that the FIS does not perform better than an Artificial Neural Network (ANN) developed for the same purpose, but useful insights about how different student attributes relate to their retention or departure may be gained from the rules that define the fuzzy model.

Acknowledgements

I would like to thank my supervisor, Professor Shamim Khan, for his excellent guidance, care and patience. I would not have made it this far without his help and encouragement. I would like to thank the School of Computer Science for providing me with an excellent learning environment throughout. Especially, I would like to thank all the faculty members and fellow students for being so helpful and inspiring to me. Furthermore, I would like to acknowledge, with much appreciation, the valuable contributions of Dr. Wayne Summers, Chair of the School of Computer Science, Dr. Mark Schmidt, Chair of the Department of Psychology and Dr. Michael Bailey, Chair of the Department of Criminal Justice, to this project. I would also like to thank fellow student Mark Plagge for helping me with his excellent Artificial Neural Network (ANN) models. Last, but not the least, I am grateful to my family members for supporting me all the way to pursue this great venture.

Table of Contents

List of Figures

List of Tables

1. Introduction

Fuzzy logic is a superset of conventional (Boolean) logic that has been extended to handle the concept of partial truth - truth values between "completely true" and "completely false" (Horstkotte, 1994). It was introduced by Lotfi Zadeh at U.C. Berkeley in the 1960s. Real situations are very often not crisp and deterministic, and they cannot be described precisely (Klir et al. 1995). In fuzzy logic, fuzzy sets are sets whose elements have degrees of membership. Zadeh, L. A. (1965) described that such a set is characterized by a membership (characteristic) function which assigns to each object a grade of membership ranging between zero and one. Fuzzy logic and fuzzy rule based systems, which are based on fuzzy logic, provide a methodology for reasoning that can handle imprecision in rules and assertions expressed by human experts (Khan, 2011). A fuzzy inference system (FIS) is a system that uses fuzzy set theory to map inputs (features in the case of fuzzy classification) to outputs (classes in the case of fuzzy classification) (Knapp, 2004). A fuzzy inference system employing fuzzy if then rules is able to model the qualitative aspects of human expertise and reasoning processes without employing precise quantitative analyses (Khoo & Zhai, 2001; Tsaganou et al., 2002; San Pedro and Burstein, 2003; Yang et al., 2005).

High student dropout rates in colleges and universities in the United States has long been a problem. According to a report released by the National Center for Public Policy and Higher Education, a low rate of college completion is a key concern in American higher education. According to ACT (the college testing service), the national average freshmen retention rate is 65.7%. From 2005 to 2010, this rate at CSU was 71% on average. Colleges and universities across the country, including CSU, are investigating student dropout rates in order to address the overall problem of student Retention, Progression and Graduation (RPG) more effectively. The

main aim of this research project was to build a fuzzy inference based model using a hybrid knowledge extraction process to predict how likely each freshman student will be to drop their program of study at the end of their first semester. CSU University Information and Technology Services (UITS), has student RPG data dating back to 1998. This historical data was utilized to develop and evaluate the fuzzy rule-based inferencing system.

Knowledge extraction for the system was performed using a top down (symbolic) as well as a bottom-up (data-based) approach. In the top-down approach, rules for the fuzzy model were derived using the traditional knowledge extraction process involving domain expert interviews. Several persons in charge of university departments that have relatively low retention rates were interviewed to identify parameters that are significant determinants of student success. Fuzzy-rules designed using this knowledge were weighted appropriately to reflect their level of significance. In the data-based second phase of fuzzy rule derivation, a feed forward (White, H., 1989) Artificial Neural Network (ANN) already trained using the student data was subjected to weight analysis to derive additional rules for the fuzzy rule base, as well as for adjusting the significance of all rules. The data provided by UITS was also analyzed as part of the bottom-up approach for building the FIS.

2. Related Work

Terenzini et al. (1980) published a paper that describes the results of the replication of a study (Tinto, 1975). They investigated the predictive validity of a 34-item instrument designed to assess the fundamental constructs of Tinto's model of college student attrition. Design, variables, and analytical procedures virtually identical to those of the original study (done at a large independent university) were used, and this research was conducted at a large public university. The five-factor structure, found in the original study was used for underlying the 34 items. It was replicated almost exactly. The five factors described were (a) background characteristics (i.e. Family background, individual attributes, precollege schooling); (b) initial commitments (i.e. Commitment to the goal of college graduation and commitment to the institution) (c) academic and social integration; (d) subsequent goal and institutional commitments; and (e) withdrawal decisions. As in the earlier work, the Institutional and Goal Commitment Scale (Pascarella, E. T., & Terenzini, P. T., 1979) was a significant predictor of attendance behavior even after controlling for a variety of students' precollege characteristics. Potential institutional differences in faculty members' influence on retention were identified. A cross-validation classification procedure suggests the five factors are reasonably stable predictors of attrition.

Mehra, N. (1973) did a study of retention and withdrawal of university students. The objective of this study was to do a preliminary investigation into the nature and extent of student dropout problems at the University of Alberta. To this end, the academic achievements of the class of 1964 were traced term by term over a period of six years. The following areas were examined in this study: (1) A quantitative general description of relative proportions of students who graduate, those who withdraw voluntarily, and those who are asked to withdraw due to poor academic performance, (2) an examination and identification of correlates of students' staying

vs. dropping out, and (3) detection and isolation of primary predictors of the criterion variable, graduation vs. dropping out. The study demonstrates that: (1) dropping out of a university is a very complex phenomenon and a better and firmer understanding of this phenomenon would require a deeper investigation, and (2) diversity within the dropout group is a reality, and to combine all dropouts into a single category is an oversimplification of the problem.

Yusof et al. (2012) had a publication on a concise fuzzy rule base to reason about student performance based on the rough-fuzzy approach (Chen, Z., 1999). Although fuzzy inference system is a potential technique to reason about students' performance, as well as to present their knowledge status (Nedic et al., 2002; Xu et al., 2002; Kosba et al. 2003), it is a challenge when more than one factors are involved in determining their performance or knowledge status (Yusof et. al, 2009). Hence, reasoning about students' performance for multiple factors is difficult. This issue is critical considering that the human experts' knowledge is insufficient to analyze all possible conditions as the information gained is always incomplete, inconsistent, and vague. Their publication presents the proposed rough-fuzzy approach to determine important attributes and refines a fuzzy rule base into a concise fuzzy rule base.

Plagge (2012) investigated the use of ANNs to predict first year student retention rates. This work expands on previous attempts to predict student outcomes using machine-learning techniques. Using a large data set provided by Columbus State University's Information Technology department, ANNs were used to analyze incoming first-year traditional freshmen students' data over a period from 2005 to 2011. Using several different network designs, the students' data were analyzed, and a basic predictive network was devised. The overall accuracy was high when the data included the first semester grades of students. Once the dataset excluded student grades for the first semester, the overall accuracy dropped significantly. Using different

network designs, more complex learning algorithms, and better training strategies, the prediction accuracy rate for a student's return approached 75% overall.

3. Methodology and Implementation

The hybrid technique used in this experiment involves both a top-down and a bottom-up approach. In this chapter the hybrid technique and the implementation is explained. In order to create a fuzzy inference based model to detect the students who are likely to drop out, we needed to create rules for the Fuzzy Inference System (FIS). The hybrid technique for deriving these rules was divided into three parts: 1) Knowledge acquisition from domain experts; 2) Weight analysis of an artificial neural network; 3) Data and statistical analysis. These three methods were applied in combination during this study.

3.1 Knowledge Acquisition from Domain Experts

The domain experts' opinion for getting the indicators played a crucial role in deriving rules for the fuzzy system. Chairs from several departments at CSU were selected as domain experts. The data provided by UITS has the number of dropouts in several departments at CSU. We selected the chairs of departments that had a significant number of dropouts. We interviewed three departmental heads as part of the top-down approach.

The questionnaire for interview had nine questions. These questions were selected in accordance with student attributes available in the data provided by UITS. The last question was kept as an open-ended one. Based on the responses we analyzed how much the domain experts agreed or disagreed about several possible causes for student dropouts. Table 1 shows the interview questions used (the edited versions of responses given by the domain experts can be found in the Appendix section). All but the last question were aimed at seeking opinions on specific conclusions derived earlier from statistical data analysis and ANN weight analysis. The last question was open-ended and attempted to reveal any additional factors not identified previously.

Table 1. Interview Questions

Q1. Over the years, more female students dropped out of their study programs than their male counterparts. Do you think that female students are more likely to drop out? If yes, what can be the possible reasons?
Q2. Do you think that compared with out-of-state students, in-state students are less likely to drop out?
Q3. Does financial aid play a positive role towards student retention? Are students who receive financial aid less likely to drop out?
Q4. Do you agree with the notion that students failing in core subjects are more likely to drop out?
Q5. Does parents' education level play a role towards student retention? Do you agree with the notion that students with college-educated parents are less likely to drop out?
Q6. In the past, students with unmet financial need had higher dropout rates. Do you think unmet financial need can cause a student to drop out?
Q7. Is high school GPA score a factor that positively correlates to retention?
Q8. Do you think that part-time students are more likely to drop out than full-time students?
Q9. What in your opinion are the three most significant factors influencing student retention rates in your department?

In the next step, these responses were analyzed to derive rules for the fuzzy inference system rule-base. Table 2 shows the summarized domain expert responses to the first eight questions.

Table 2. Summarized responses from domain experts

Q1. Over the years more female students dropped out of their study programs than their male counterparts. Do you think that female students are more likely to drop out? If yes, what can be the possible reasons?		
Expert 1 - Response	Expert 2 - Response	Expert 3 - Response
Agree	Disagree	Disagree
Q2. Do you think that compared with out-of-state students, in-state students are less likely to drop out?		
Expert 1 - Response	Expert 2 - Response	Expert 3 - Response
Disagree	Disagree	Agree
Q3. Does financial aid play a positive role towards student retention? Are students who receive financial aid less likely to drop out?		
Expert 1 - Response	Expert 2 - Response	Expert 3 - Response
Agree	Strongly agree	Agree
Q4. Do you agree with the notion that students failing in core subjects are more likely to drop out?		
Expert 1 - Response	Expert 2 - Response	Expert 3 - Response
Strongly agree	Strongly agree	Agree
Q5. Does parents' education level play a role towards student retention? Do you agree with the notion that students with college-educated parents are less likely to drop out?		
Expert 1 - Response	Expert 2 - Response	Expert 3 - Response
Neutral	Strongly agree	Agree
Q6. In the past, students with unmet financial need had higher dropout rates. Do you think unmet financial need can cause a student to drop out?		
Expert 1 - Response	Expert 2 - Response	Expert 3 - Response
Strongly agree	Strongly agree	Strongly agree
Q7. Is high school GPA score a factor that positively correlates to retention?		
Expert 1 - Response	Expert 2 - Response	Expert 3 - Response
Agree	Agree	Disagree
Q8. Do you think that part-time students are more likely to drop out than full-time students?		
Expert 1 - Response	Expert 2 - Response	Expert 3 - Response
Agree	Agree	Disagree

At this point, it was clear that on some of the issues the domain experts were in agreement. These were obvious choices for inclusion in the rules. For example, the response to question 6 made it

relatively easy to form a rule since all domain experts strongly agreed to the fact that unmet financial need can cause a student to drop out. So we can create a rule as follows:

If unmet financial need is high then dropout likelihood is high

Similarly, we can consider responses to question no. 3, where one of the three domain experts strongly agreed, and the other experts agreed, to the fact that financial aid plays a positive role towards student retention. This led to the formation of the following rule:

If received financial aid is high then dropout likelihood is low

For this study we have used "financial need difference" instead of "financial aid" as an input. The reason was simply because the available data provided by UITS includes this field. Financial aid difference is the difference between the amount required for educational expenses and the financial aid received by a student. So we used the following two rules instead of the ones mentioned above:

If financial need difference is high then dropout likelihood is high

If financial need difference is low then dropout likelihood is low

But when the domain experts were divided in their opinions, it became difficult to derive rules like the ones mentioned above. For example, if we look at the response of the first question, we can see that two out of the three domain experts disagreed about the notion that female students are more like to drop out. The tool (Fuzzy toolbox for Matlab) allows us to adjust the rule execution weights. In such cases, weights of less than 1.0 were associated with the rules to reduce their contribution in the reasoning process.

A fuzzy rule's inclusion in the rule base also depended on the other two methodologies. If those bottom-up data-based approaches had supported the notion, the corresponding fuzzy rule was kept in the rule base. The following table (Table 3) shows rules that were derived by interviewing the three departmental heads at CSU. The rules here are listed in descending order from the most important to the least important one (determined by rule confidence expressed as weights).

Table 3. List of rules derived from domain experts' opinions

Rule no.	Rule	Reason for choosing rule
1	*If estimated family contribution is low then dropout likelihood is high.*	Rule based on response to question 6. All domain experts strongly agreed.
2	*If financial need difference is high then dropout likelihood is high.*	Rule based on responses to question 6. All experts strongly agreed.
3	*If students fail in core courses then dropout likelihood is high.*	Rule based on responses to question 4. All experts agreed.
4	*If father's highest education level is at least college or mother's education level is at least college then dropout likelihood is low.*	Rule based on response to question 5. All experts agreed.
5	*If high school GPA is high then dropout likelihood is low.*	Rule based on responses to question 7. Two experts agreed, one disagreed. So this rule may be implemented lower execution weight.
6	*If high school GPA is low then dropout likelihood is high.*	Rule based on the responses to question 7. All experts agreed.
7	*If student's status is out-of-state then dropout likelihood is high.*	Rule based on responses to question 2. Two experts agreed, one disagreed. Rule assigned lower weight.
8	*If gender is female then dropout likelihood is high.*	Rule based on responses to question 1. Two experts agreed, one disagreed. Rule assigned lower weight.

3.2 Rule Extraction from Artificial Neural Network

The second approach employed in the hybrid knowledge extraction process was the bottom-up or data-based approach. It involved the extraction of fuzzy rules from an ANN that had been already created in a separate project by Plagge (2012). This feedforward ANN was trained to predict if a student is likely to return to the University after his or her first year.

In a feedforward ANN, information moves from the input neurons into the neurons of the hidden layer and then into the output layer neurons. The feedforward ANN in Figure 1 has three input layer neurons and three neurons in the hidden layer, and one neuron in the output layer. Each of the input neurons is connected to the neurons of the next layer with different weights. The weight determines the strength of the signal being transmitted from one neuron to another. For example, W_{14} represents the weight from input neuron N1 to hidden layer neuron N4. During training, these weights are adjusted to reduce error in the classification of the input patterns by the ANN. A successfully trained ANN is said to have learned the functional relationship between its input and the corresponding expected output.

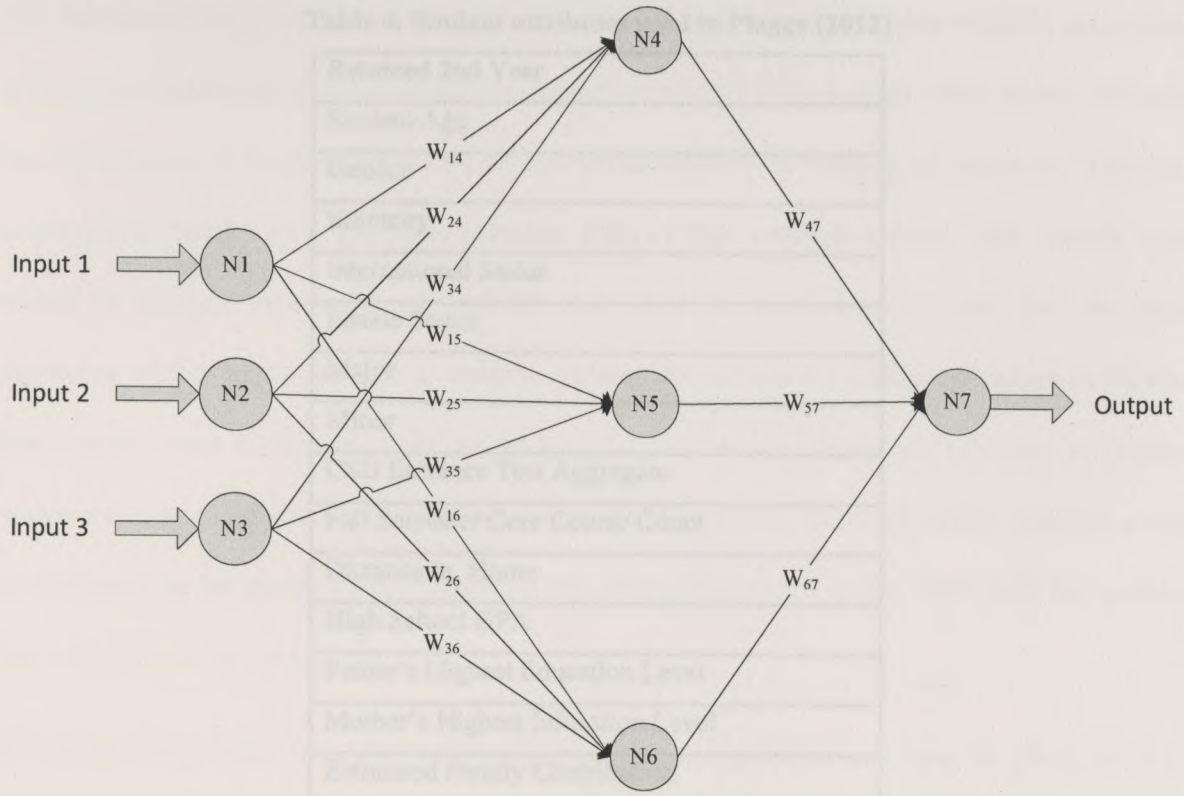

Figure 1. A feedforward ANN

The data used in Plagge (2012) was the same as the data used in this study for training and testing the fuzzy inference system, and was provided by the UITS. It consisted of attributes of CSU freshmen entering in the fall semester over six years during the period 2005 - 2010. There were a large number of variables for each student record in that dataset. In order to train the ANN effectively, any attributes regarded as irrelevant were removed. The resulting reduced dataset had 16 variables per student as shown in Table 4.

Table 4. Student attributes used in Plagge (2012)

Returned 2nd Year
Student Age
Gender
Ethnicity
International Status
Instate Status
Major
Minor
CSU Entrance Test Aggregate
Fall Semester Core Course Count
Distance to Home
High School GPA
Father's Highest Education Level
Mother's Highest Education Level
Estimated Family Contribution
Financial Need Difference

Plagge (2012) built and trained several ANN models, of which, the most successful one had an overall accuracy of 76.09% as shown in the confusion matrix (in Table 5). The confusion matrix allows us to visualize the performance of the ANN by giving both correct (0 classified as 0, 1 classified as 1) and incorrect classifications. Here 0 denotes students who dropped out after their first semester and 1 represents students who returned to continue their studies.

Table 5. ANN Confusion Matrix

Actual \ Predicted	**0**	**1**	**No. of records**	**% correct**
0	764	1105	1869	40.88%
1	319	3767	4086	92.19%
No. of records	1083	4872	5955	76.09%

One important thing to notice here is that this confusion matrix shows only 40.88% correctness of dropout predictions whereas retention prediction was 92.19% correct. This clearly indicates that the learning of dropouts by the ANN was not as good as its learning of retentions. This may be explained by the fact that approximately 69% of the available student data records were related to students who returned, and the data used to train the ANN also had the same proportion of the two categories. In order to address this imbalance, additional copies of dropout data records were made and added to the original data set to bring the retention to dropout record ratio up to almost 1 (3738 dropout and 4085 retention records) before training a new model based on the design of the existing ANN. The performance of the ANN with this updated data set is shown in Table 6.

To extract fuzzy rules from the ANN, we followed an algorithm proposed by Muslimi et al. (2008). This approach falls in the category of decompositional algorithm (Andrews, R. et al. 1995), where hidden or output layer nodes are analyzed individually. In Figure 1, we can see a two-layer feedforward ANN. Most ANN rule extraction algorithms use the maximum weight linking a neuron to neurons in the following layer to extract fuzzy rules. It assumes that how an input variable affects the activation of an output neuron depends only on the maximum inter-neuron weight associated with that variable, and not on the minimum weight. This incorrect assumption causes antecedents that can be pruned to sometimes escape pruning, making the rules less general, and consequently diminishing the accuracy of the fuzzy inference system implementing the extracted rules. The extraction algorithm proposed by Muslimi et al. (2008) claims to overcome this problem.

In this process of rule extraction, each variable applied as input to the ANN is decomposed into two or more binary variables. For instance, the input variable Gender has 2 possible values: male

and female. So we used 2 input neurons labeled Gender-Male and Gender-Female instead of just one variable Gender (see Figure 2). It should be noted that the binary input variables within each input parameter group are dependent on other variable(s) in the group. For example, if Gender is Gender – Male (value set to 1), then Gender – Female will be set to value 0.

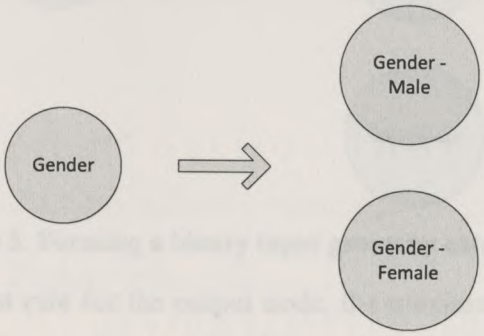

Figure 2. Forming a binary input group by all possible values

The remaining 15 input variables were similarly split into binary input groups. This was done either by breaking them up into all possible values or into all possible categories. In the above example, we have broken the input Gender into the two possible values of it to form an input group. But for variables such as High-school-GPA, it was more sensible to break it down into categories HSGPAHigh, HSGPALow and HSGPAModerate (Figure 3). Eventually there were 140 binary input variables belonging to 15 binary valued groups in the revised ANN created to extract fuzzy rules for our experiment.

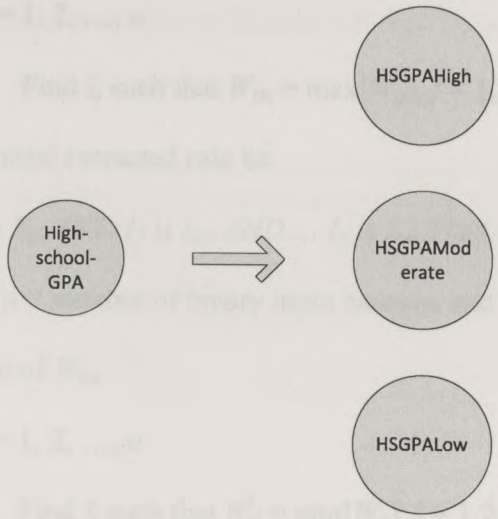

Figure 3. Forming a binary input group by categories

To extract the most dominant rule for the output node, the maximum weight W_{im} of each input parameter I_i is determined. Here i ranges in value from 1 to n (number of input neurons). For each input neuron, we added up all the weights linking it to the next layer neurons. Similarly the minimum weight W_{il} for each binary input was calculated. The input neurons are then sorted in ascending order by the absolute difference of W_{im} and W_{il}. In the last step, the algorithm prunes the binary input neurons, starting with the smallest absolute difference of W_{im} and W_{il}, so long as the neuron remains activated if its maximum-weight binary input is off and the minimum-weight binary input is on. The pseudo code for extracting fuzzy rules from ANN is as follows:

For $i = 1, 2, \ldots, n$

Find I_i such that $W_{im} = \max[W_{ij}]$, $j = 1,2,3$.

Let, initial extracted rule be

If I_1 is I_{1m} AND I_2 is I_{2m} AND.... I_n is I_{nm} Then Output is O.

Here, n = number of binary input neurons and im = corresponding binary neuron of W_{im}

For $i = 1, 2, \ldots, n$

Find I_i such that $W_{il} = \min[W_{ij}]$, j = 1,2,3.

Find $d_i = |W_{im} - W_{il}|$

Sort I_i in ascending order of d_i

Let, $S = \sum_{i=1}^{n} W_{im} - B$

Here, B is the bias for ANN.

For $i = 1,2,\ldots,n$

$S = S - W_{im} + W_{il}$,

If $S<0$

Exit

Else

Remove antecedent that involves parameter I_i

A utility program for this experiment was written in Microsoft .Net platform (Figure 4). This program can generate binary input values from the available data, analyze weights and generate fuzzy rules from the imported weights using the above mentioned algorithm. Additionally, this

program is also able to generate test data for Matlab's Fuzzy Logic Tool that we used for this study.

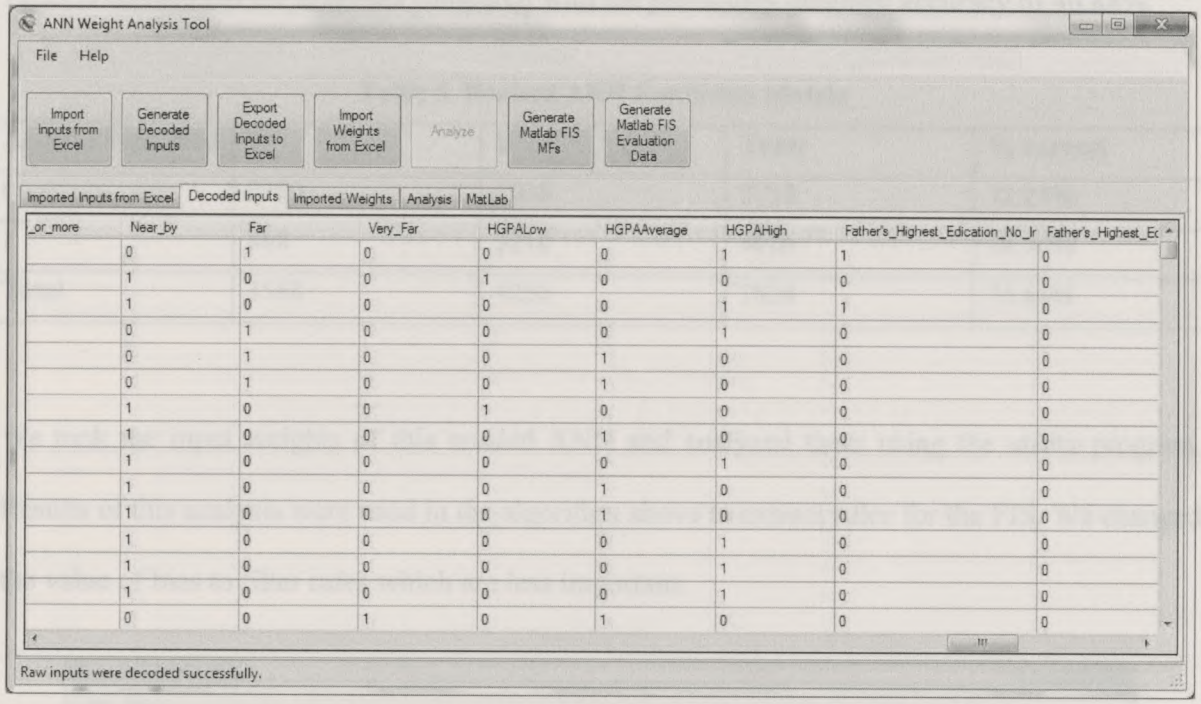

Figure 4. A screenshot of the utility program execution

This program was first used to encode all 15 inputs to binary. Those inputs were used for the revised version of the ANN. We created over 20 ANN models with different numbers of layers and number of neurons in the hidden layer(s) to search for the best possible outcome in terms of prediction accuracy. Out of these, the best one was a 4-layer feed-forward ANN (Figure 5). This ANN has 16 neurons in the first hidden layer, 64 neurons in the second hidden layer and 128 neurons in the third hidden layer.

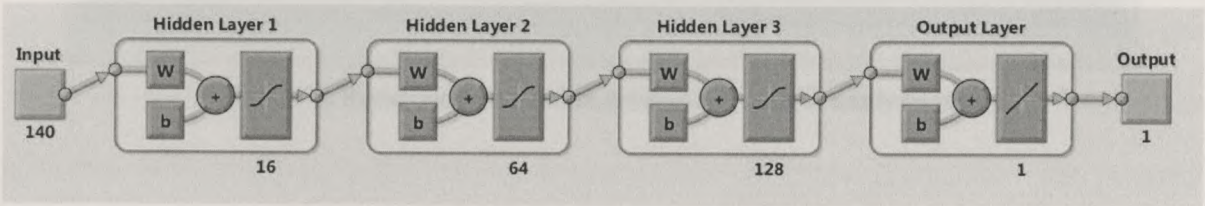

Figure 5. Revised version of ANN as displayed in Matlab

The overall prediction accuracy (see Table 6 below) of the revised ANN is almost as good as the previous ANN developed in (Plagge, 2012) (see Table 5 above), but it gives a higher accuracy of 72.23% correctness for dropouts compared with the previously obtained accuracy of 40.88%.

Table 6. Revised ANN Confusion Matrix

Actual \ Predicted	0	1	Total	% correct
0	2700	1038	3738	72.23%
1	868	3218	4086	78.76%
Total	3568	4256	7824	75.64%

We took the input weights of this revised ANN and analyzed them using the utility program. Results of this analysis were used in the algorithm above to extract rules for the FIS. We changed the value of bias to filter rules which are less important.

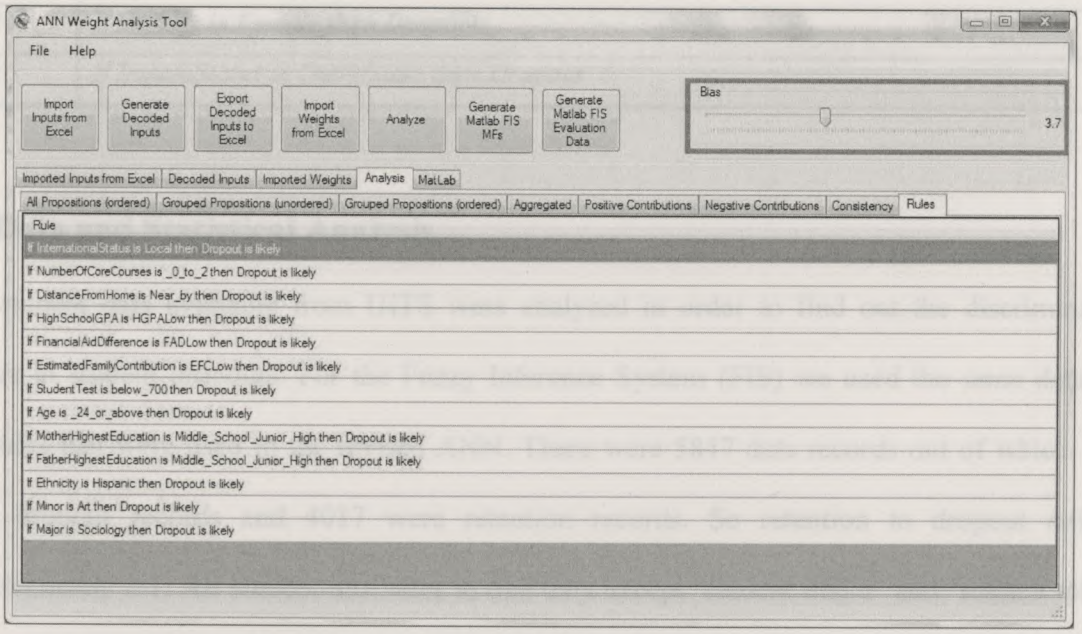

Figure 6. Fuzzy rule extraction using ANN weight analysis tool

The following table lists rules extracted from the revised ANN:

Table 7. Extracted fuzzy rules from ANN

Rule no.	Rule
1	*If Major is Sociology then Dropout*
2	*If Minor is Art then Dropout*
3	*If Ethnicity is Hispanic then Dropout*
4	*If FatherHighestEducation is MiddleSchoolJuniorHigh then Dropout*
5	*If MotherHighestEducation is MiddleSchoolJuniorHigh then Dropout*
6	*If Age is Older then Dropout*
7	*If StudentTestScore is Low then Dropout*
8	*If EstimatedFamilyContribution is Low then Dropout*
9	*If FinancialNeedDifference is Needy then Dropout*
10	*If HighSchoolGPA is Low then Dropout*
11	*If DistanceToHome is NearBy then Dropout*
12	*If NumberOfCoreCourses is Low then Dropout*
13	*If InternationalStatus is Local then Dropout*
14	*If Gender is Female then Dropout*
15	*If InstateStatus is OutOfState then Dropout*

3.3 Data and Statistical Analysis

The student data available from UITS were analyzed in order to find out the discriminatory factors in student dropouts. For the Fuzzy Inference System (FIS) we used the same data and variables that were used in the revised ANN. There were 5847 data records out of which 1829 were dropout records and 4017 were retention records. So retention to dropout ratio is approximately 2:1. All student-attributes in that data except 'student major' and 'student minor' were considered for analysis. For each attribute dropout rates and dropout percentages were calculated. If dropout percentage was low for any value or particular group, that value or group was disregarded. This is because a low percentage of student population with low percentage of

dropout has negligible impact in overall dropouts. Again, a value or group with high dropout rate that has significant student population with dropout percentage was considered to be included in the rule base. The input variable selection process has been discussed in detail in section 3.4.

The student attributes are discussed below:

Student age

From the data (Table 8 and Figure 6) it looks quite evident that most students are aged 16 to 18 years. The dropout rate (30.0%) is quite high but is not very significant when the overall dropout rate is around 30%. Rather young people (aged between 19 and 23) have a higher dropout rate of 42.3% but that contributes to only 13.7% of overall dropouts. There is only a very few students aged over 23.

Table 8. Data analysis: Student age

	Teen (16 to 18)	Young (19 to 23)	Older (24 or above)
Percentage in student population	89.5%	10.1%	0.3%
Dropouts	1572	251	7
Retentions	3663	342	12
Dropout Rates	30.0%	42.3%	36.8%
Dropout percentage	85.9%	13.7%	0.4%

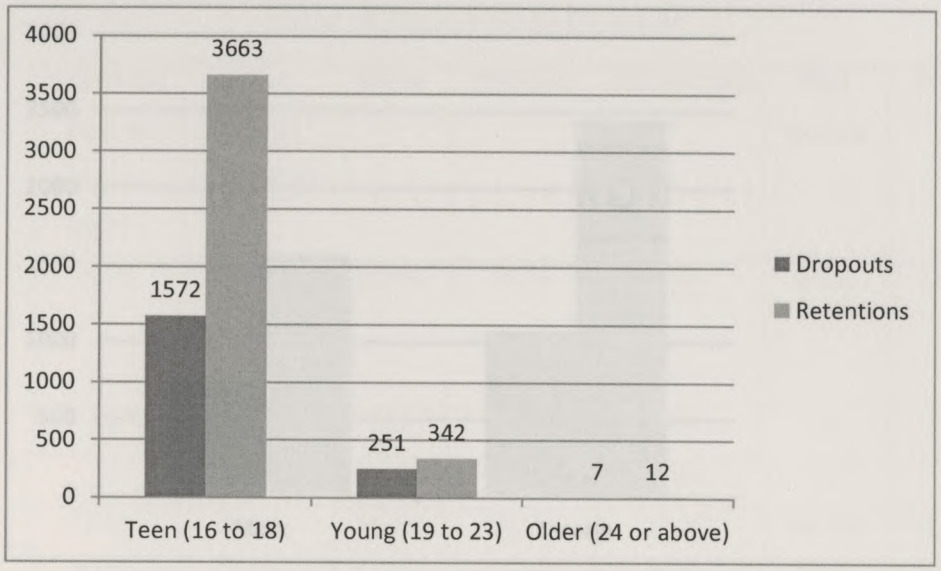

Figure 6. Data analysis: Student age

There are two fuzzy rules (see Table 21 in section 3.4) related to student age. However, there was no rule related to age group 'older' since less than 1% students are in that age group.

Gender

Both male and female have similar dropout rates but female students contribute more than their male counterparts (Table 9 and Figure 7).

Table 9. Data analysis: Gender

	Male	Female
Percentage in student population	39.7%	60.3%
Dropouts	756	1074
Retentions	1567	2450
Dropout Rates	32.5%	30.5%
Dropout percentage	41.3%	58.7%

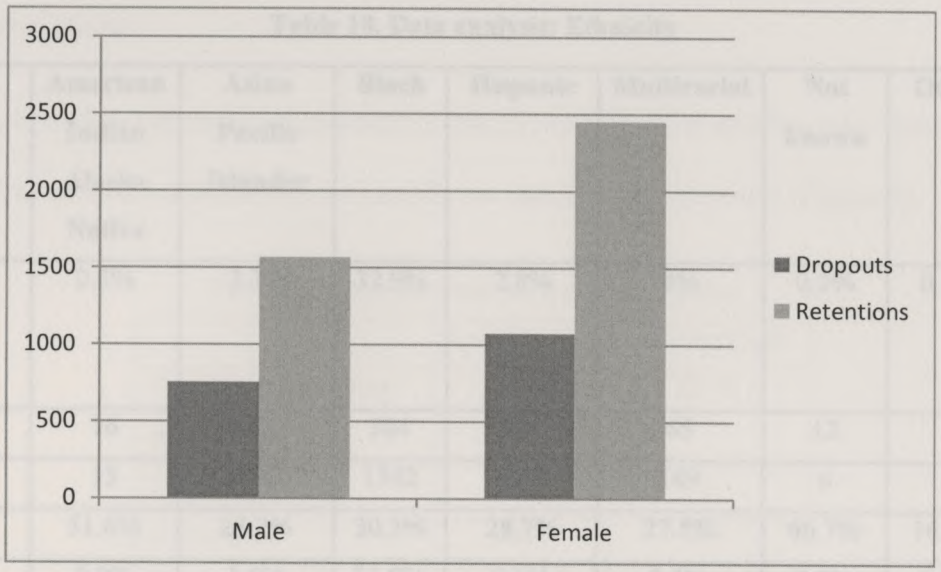

Figure 7. Data analysis: Gender

There was no rule related to gender since it is not a discriminatory student attribute for dropouts.

Ethnicity

Figure 8 clearly shows that the number of Hispanic or other students are not very significant compared to students who are either white or African American (black). Both white and black has similar dropout rates.

Table 10. Data analysis: Ethnicity

	American Indian Alaska Native	Asian Pacific Islander	Black	Hispanic	Multiracial	Not known	Other	White
Percentage in student population	0.5%	2.3%	32.9%	2.8%	4%	0.3%	0.1%	57%
Dropouts	16	32	584	47	65	12	1	1073
Retentions	15	106	1342	117	169	6	5	2257
Dropout Rates	51.6%	23.2%	30.3%	28.7%	27.8%	66.7%	16.7%	32.2%
Dropout percentage	0.9%	1.8%	31.9%	2.6%	3.6%	0.7%	0.1%	58.6%

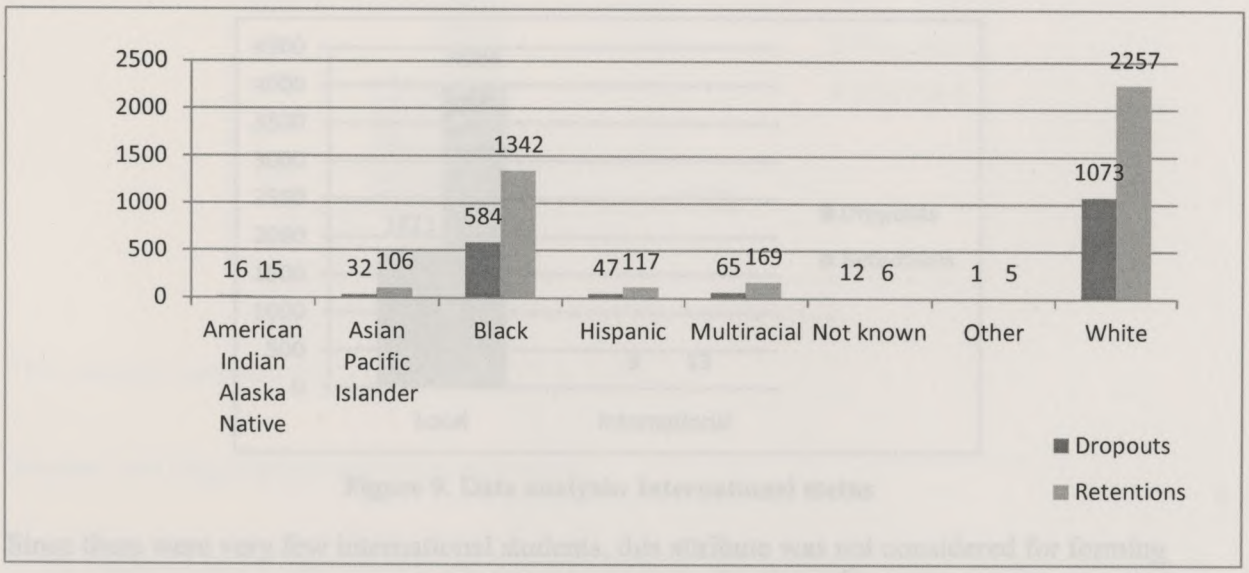

Figure 8. Data analysis: Ethnicity

From the analysis above we can see that none of the ethnicity group with significant student population is proved to be discriminatory factor for dropouts.

International status

There are only a few international students (Table 11) at CSU. The number is so small compared to the number of local students that the bars do not appear in the bar chart (Figure 9).

Table 11. Data analysis: International status

	Local	International
Percentage in student population	99.6%	0.4%
Dropouts	1821	9
Retentions	4004	13
Dropout Rates	31.3%	40.9%
Dropout percentage	99.5%	0.5%

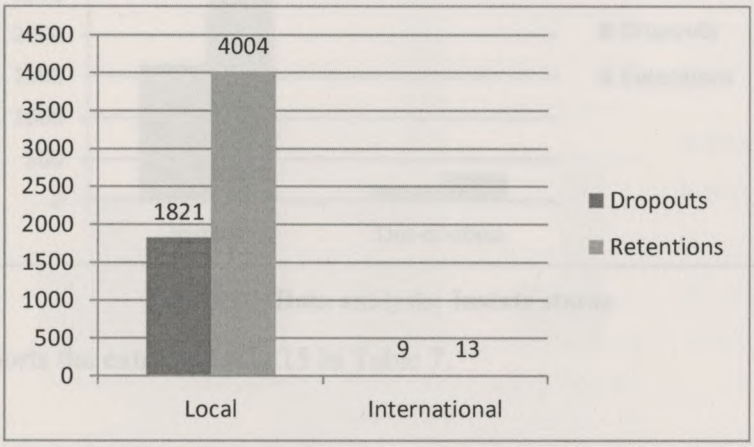

Figure 9. Data analysis: International status

Since there were very few international students, this attribute was not considered for forming rules.

Instate status

About 10% of the students are out-of-state students (Table 12). Dropout rates are higher for out-of-state students than instate students.

Table 12. Data analysis: Instate status

	Instate	Out-of-state
Percentage in student population	91%	9%
Dropouts	1645	185
Retentions	3676	341
Dropout Rates	30.9%	35.2%
Dropout percentage	89.9%	10.1%

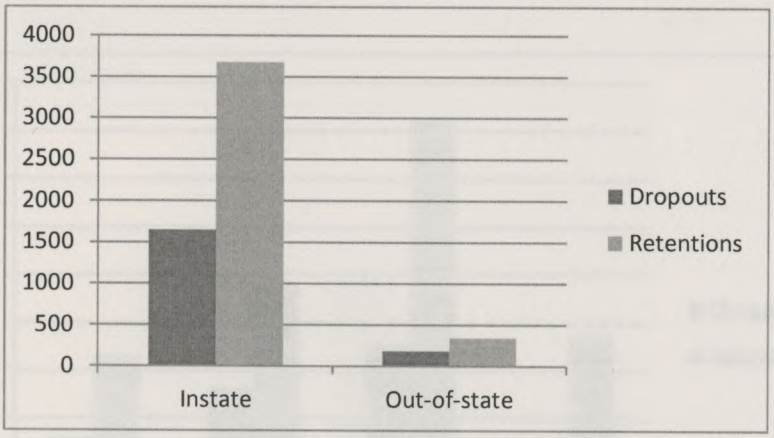

Figure 10. Data analysis: Instate status

This analysis supports the extracted rule 15 in Table 7.

Student test (aggregated entrance test score)

From Table 13 and Figure 11 we can see that dropout rate increases when student test score is low and decreases when test score is high. About 17% of the data are unknown.

Table 13. Data analysis: Student test score

	Unknown	Low (<900)	Moderate (800 - 1100)	High(1100 +)
Percentage in student population	16.8%	25.4%	40.3%	17.5%
Dropouts	319	526	713	272
Retentions	664	959	1643	751
Dropout Rates	32.5%	35.4%	30.3%	26.6%
Dropout percentage	17.4%	28.7%	39%	14.9%

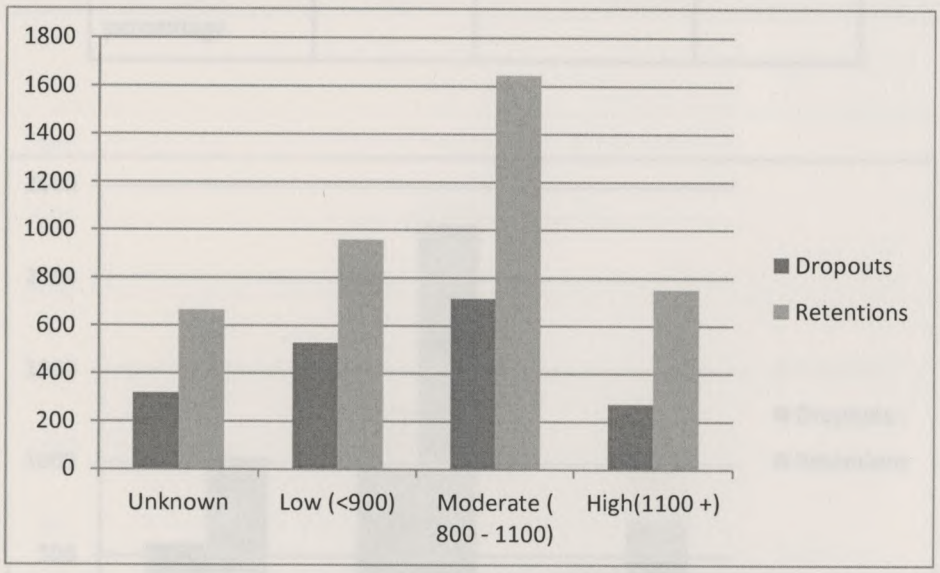

Figure 11. Data analysis: Student test score

Student test score was considered for creating fuzzy rules as the analysis above supports the extracted rule 7 in Table 7. Also, two more rules were incorporated in the FIS with different confidence levels.

Course load

As per the data in Table 14 dropout rate is higher for those taking two or fewer courses.

Table 14. Data analysis: Course load

	Low (0-2)	Moderate (3-4)	High (5+)
Percentage in student population	27.2%	55.9%	16.9%
Dropouts	563	978	289
Retentions	1028	2292	697
Dropout Rates	35.4%	29.9%	29.3%
Dropout percentage	30.8%	53.4%	15.8%

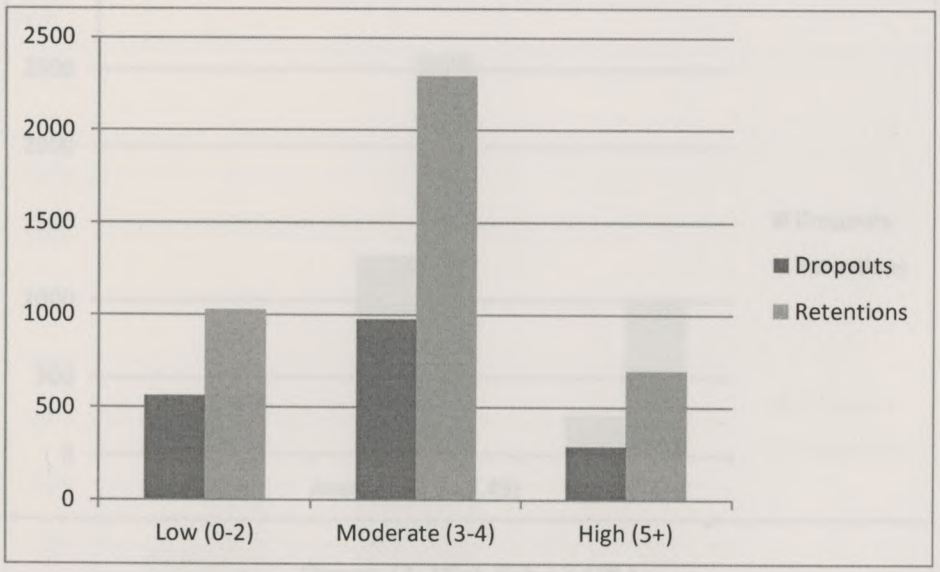

Figure 12. Data analysis: Course load

This leads us to form a rule with very high confidence level that relates to students with low course loads. Two other rules with lower confidence level was also included which relate to course load.

High School GPA

From Table 15 and Figure 13 we see that dropout rate is very high for students who have low high school GPA. On the other hand dropout rate is significantly low for students who have high school GPA over 3.5.

Table 15. Data analysis: High School GPA

	Low (<2.5)	Average (2.5 - 3.49)	High (3.5 +)
Percentage in student population	11.4%	66.7%	21.9%
Dropouts	278	1293	259
Retentions	391	2606	1020
Dropout Rates	41.6%	33.2%	20.3%
Dropout Percentage	15.2%	70.7%	14.2%

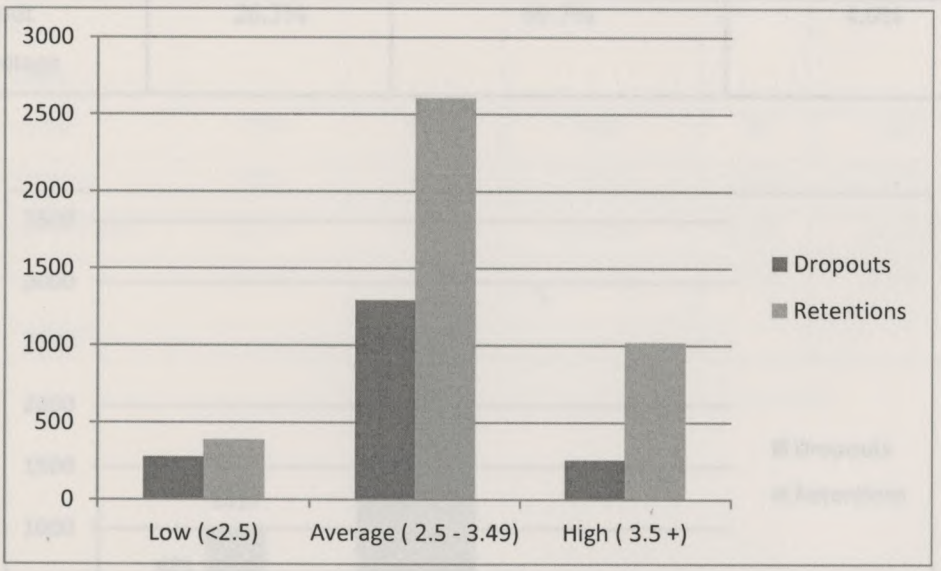

Figure 13. High School GPA

The analysis above supports rule 5 in Table 3 and rule 10 in Table 7.

Distance to home

From Table 16 we can see that most students' home is between 69 and 200 miles away. But surprisingly dropout rate is slightly higher for those whose home is nearby. When the distance from home is very far, the dropout rate increases to over 36%.

Table 16. Data analysis: Distance to home

	Near (< 69 miles)	Far (>69 and <200 miles)	Very Far (>200 miles)
Percentage in student population	25.7%	70.9%	3.4%
Dropouts	481	1275	74
Retentions	1017	2873	127
Dropout Rates	32.1%	30.7%	36.8%
Dropout percentage	26.3%	69.7%	4.0%

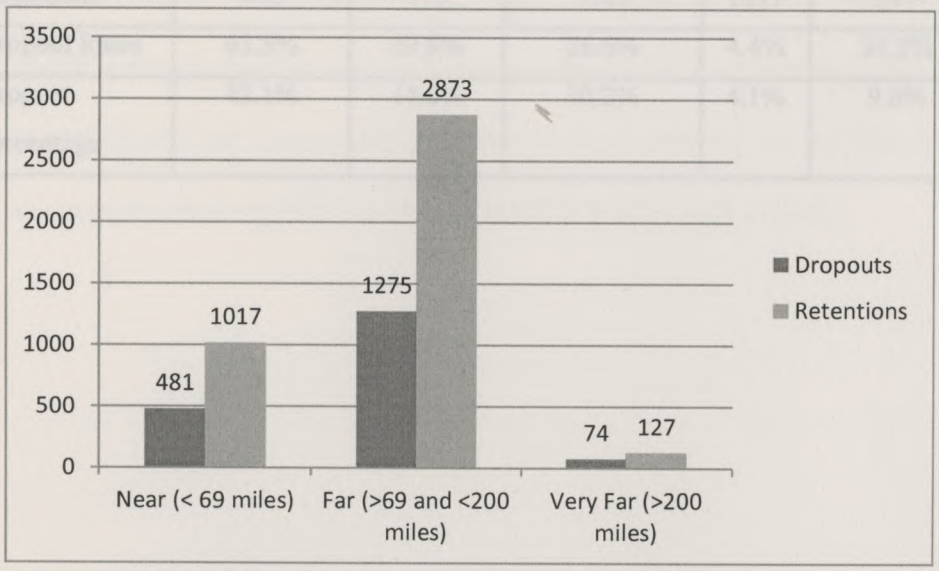

Figure 14. Data analysis: Distance to home

From the analysis we see that only the students whose distance to home is very far can be considered more likely for dropouts. But they are very few in number.

Father's highest education level

It is not very clear from Table 17 when the dropouts are higher since a lot of data fall in the category of other or unknown. About 42% dropout records have no information about father's highest education level.

Table 17. Data analysis: Father's highest education level

	No Information	Other or Unknown	Middle School Junior High	High School	College or Beyond
Percentage in student population	20.7%	6.6%	35.5%	28.9%	8.2%
Dropouts	770	268	552	75	165
Retentions	442	116	1525	1617	317
Dropout Rates	63.5%	69.8%	26.6%	4.4%	34.2%
Dropout percentage	42.1%	14.6%	30.2%	4.1%	9.0%

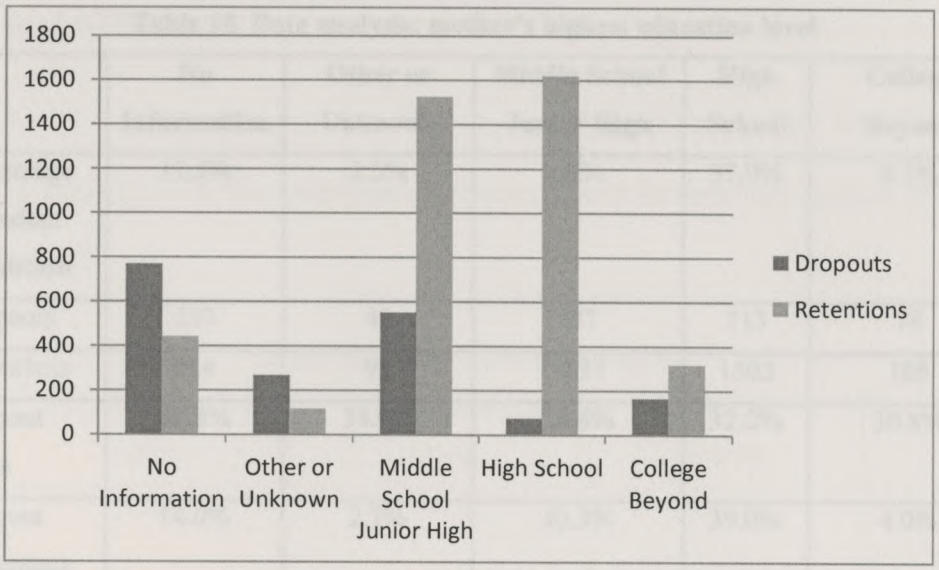

Figure 15. Data analysis: Father's highest education level

Over 20% of the student records have no information regarding father's highest education level. This leads to keep lower confidence level in any fuzzy rule related to this attribute.

Mother's highest education level

Mother's highest education level (Table 18 and Figure 16) gives us slightly better insights since fewer records fall in the category of other or unknown than father's highest education level. Only 14% dropout contributors are untraceable. Ignoring those records we can see that dropout rates are higher for students whose mother's highest education level is high school.

Table 18. Data analysis: mother's highest education level

	No Information	Other or Unknown	Middle School Junior High	High School	College Beyond
Percentage in student population	11.5%	2.5%	44%	37.9%	4.1%
Dropouts	257	49	737	713	74
Retentions	414	95	1837	1505	166
Dropout Rates	38.3%	34.0%	28.6%	32.2%	30.8%
Dropout percentage	14.0%	2.7%	40.3%	39.0%	4.0%

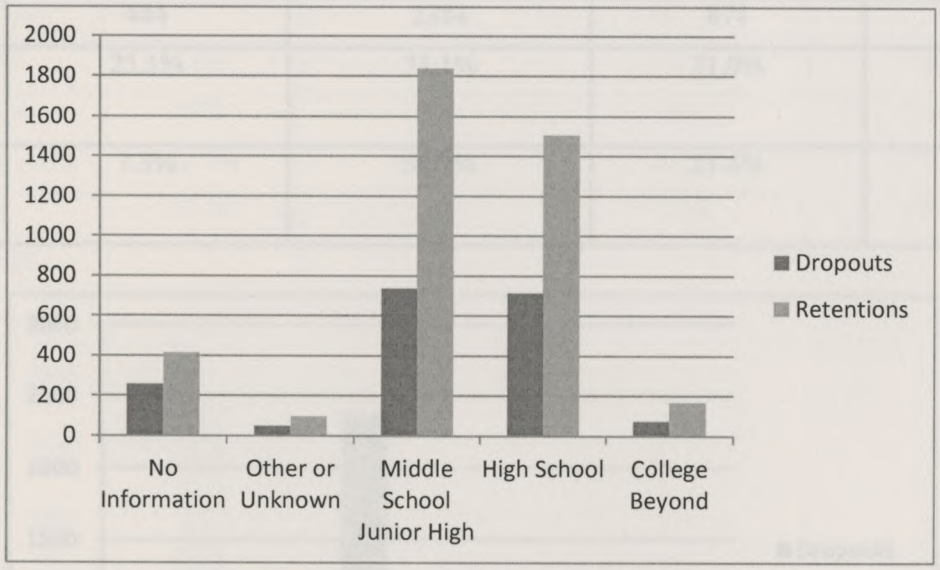

Figure 16. Data analysis: Mother's highest education level

Considering mother's highest education level for fuzzy rules, the same has been done as was done for father's education level.

Financial need difference

From Figure 17 it is very evident that dropout rate is significantly high when financial aid difference is very high. On the other hand, retention rate is the best for students who are very well off (Table 19).

Table 19. Financial need difference

	Very well off (< -3000)	Well off (-3000 to 3000)	Needy(3000+ - 10000)	Very Needy (10000+)
Percentage in student population	9.9%	59.1%	21.7%	9.3%
Dropouts	133	1074	392	232
Retentions	444	2384	874	315
Dropout Rates	23.1%	31.1%	31.0%	42.4%
Dropout percentage	7.3%	58.7%	21.4%	12.7%

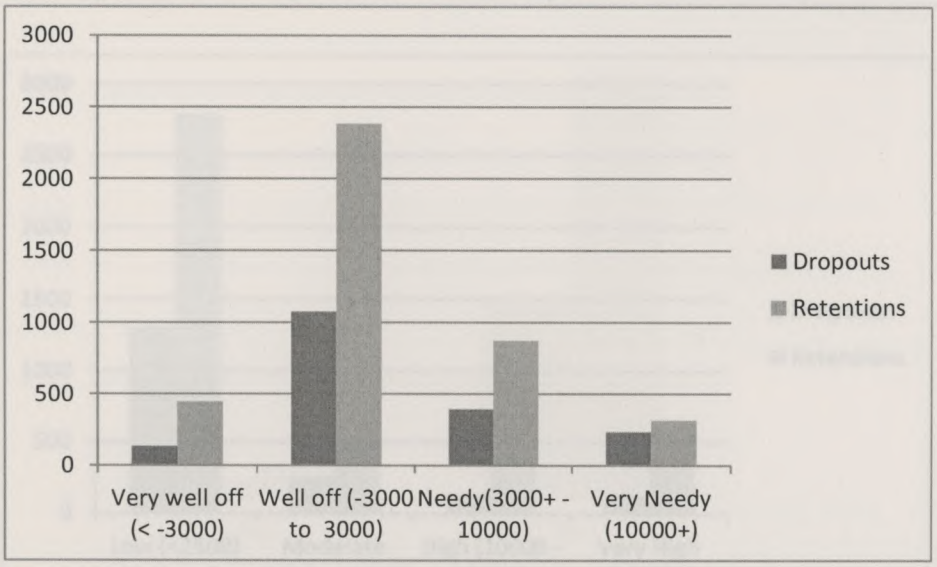

Figure 17. Data analysis: Financial need difference

The last column in Table 19 supports the notion that dropout rate is very high for students who are very needy.

Estimated Family Contribution (EFC)

From Table 20 we can see that retention rate is high for students with higher EFC. From Figure 18 it is quite evident that most students have low estimated family contributions.

Table 20. Data analysis: Estimate Family contribution (EFC)

	Low (<2500)	Moderate (2500 - 10000)	High (10000 - 20000)	Very High (20000+)
Percentage in student population	70.1%	13.1%	8.5%	8.4%
Dropouts	1308	256	133	133
Retentions	2791	508	362	356
Dropout Rates	31.9%	33.5%	26.9%	27.2%
Dropout percentage	77.1%	15.1%	7.8%	7.8%

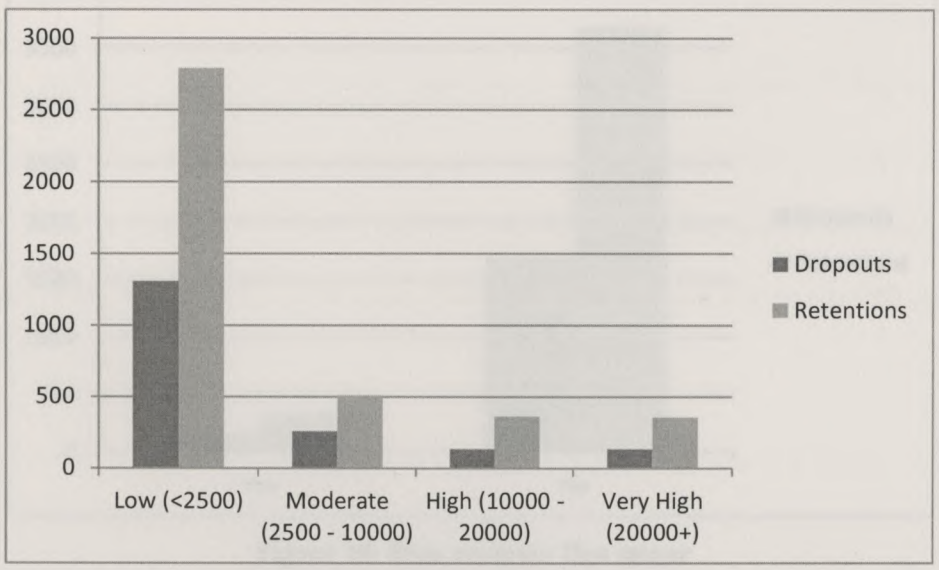

Figure 18. Data analysis: Estimated Family Contribution (EFC)

The analysis above supports rule 1 in Table 3 rule 9 in Table 7.

Has minor

From Figure 19 we can see that only a very few students have minor. Table 21 indicates that there is not much of a difference between dropout rates of those who have minor and dropout rates of those who have no minor.

Table 21. Data analysis: Has minor

	Yes	No
Percentage in student population	8.1%	91.9%
Dropouts	144	1686
Retentions	332	3685
Dropout Rates	30.3%	31.4%
Dropout percentage	7.9%	92.1%

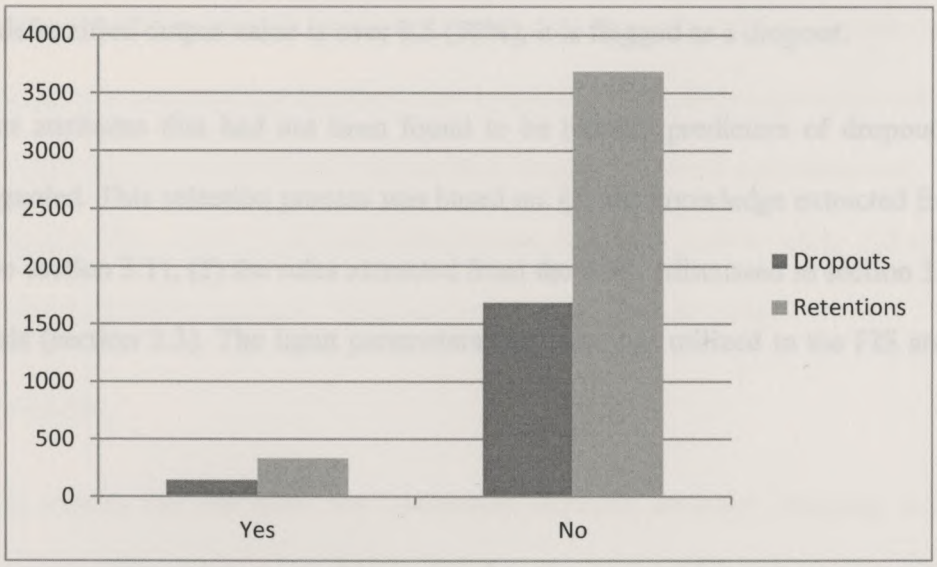

Figure 19. Data analysis: Has minor

From Table 20 we can see that this attribute does not support to form a rule with high confidence level.

3.4 Creating the Fuzzy inference system

There is no rule of thumb about creating a Fuzzy Inference System, which can ensure the best outcome. It has to be tuned on a trial and error basis. We chose the rules based on the available information discussed in the last three sections. The fuzzy sets were fine-tuned after several trials.

We followed Mamdani's fuzzy inference method (Mamdani, E. H., 1974) for building the Fuzzy Inference System (FIS). In that methodology, after the aggregation process, there is a fuzzy set for each output variable that needs defuzzification (MathWorks, 2012). The output variable has a range from zero to one (Figure 20). The centroid method was used for defuzzification. So, the output depends on the aggregated clipped regions contributed by all the output fuzzy sets for output variable DropoutChance (name used in in Matlab for output variable dropout likelihood). When the defuzzified output value is over 0.5 (50%), it is flagged as a dropout.

The student attributes that had not been found to be reliable predictors of dropout likelihood were disregarded. This selection process was based on: (1) the knowledge extracted from domain experts (see section 3.1); (2) the rules extracted from the ANN (discussed in section 3.2); and (3) data analysis (section 3.3). The input parameters that were not utilized in the FIS are discussed below:

Gender:

From Table 3 (section 3.1) we find that there is a less important rule (rule 8) at the bottom of the table. In Table 7 (section 3.2) there was an extracted rule related to Gender (rule 14). This was also listed as one of the least important. When we did the data analysis we found that both male and female students showed similar dropout rates (Table 9).

International Status:

There was no rule formed by the knowledge extraction which relates to the international status of students (Table 3). The fuzzy rule extraction from ANN algorithm lists an unimportant rule at the bottom of Table 7 (rule 13) that is related to international status of students. If we look at the data in Table 11, over 99% of the contributors were dropout records of the local students.

Ethnicity:

The domain experts did not infer anything related to ethnicity of the students. So we do not see any rule is listed in Table 3 that is based on ethnicity. However, in Table 7 we can see that there is an important rule suggested by the fuzzy rule extraction from ANN algorithm (rule 3). But we left this out after data analysis. Table 7 shows that Hispanic students contribute only 2.6% to the overall dropouts. The most significant contributors were the native white (58.6%) and the African American (31.9%) students. They had very similar dropout rates.

Distance to home:

The domain experts did not infer any correlation between students' dropping out and there distance to home. Table 7 lists a rule that is related to the distance from home (rule 11). But if we look at the data in Table 15 it is not very clear that there is a significant difference between

dropout rates of students whose distance to home is nearby and those whose distance to home is far. Although dropout rate increases if the distance is very far, the contribution (about 4%) in dropout records is too low to consider it for a rule.

Major:

There were records for over 50 different majors. Although the fuzzy rule extraction algorithm lists an important rule (rule 1 in Table 7), we left it out due to the fact that we would need to define Boolean variables for each major. Also, categorization all majors was not possible.

Minor:

Like major, there were records for over 30 minors. We can find an important rule listed in Table 7 (rule 2). But we left it out for the same reason as for majors. There are only 7.9% students contributing to dropouts who have minors. So, it was decided to keep a Boolean variable 'HasMinor' that simplifies the task of creating fuzzy sets for minors.

In the next step, the fuzzy sets were created for all the other input variables (including 'HasMinor') and the output variable (DropoutChance). For all of these variables adjustments to the fuzzy sets were made in order to improve the accuracy of the results for detecting student dropouts. The designs of each of these are discussed below:

Output variable (DropoutChance):

The output variable has a range from zero to one (Figure 20). There are 4 fuzzy sets: Low, Moderate, High and VeryHigh for the output variable. All of them were Gaussian functions.

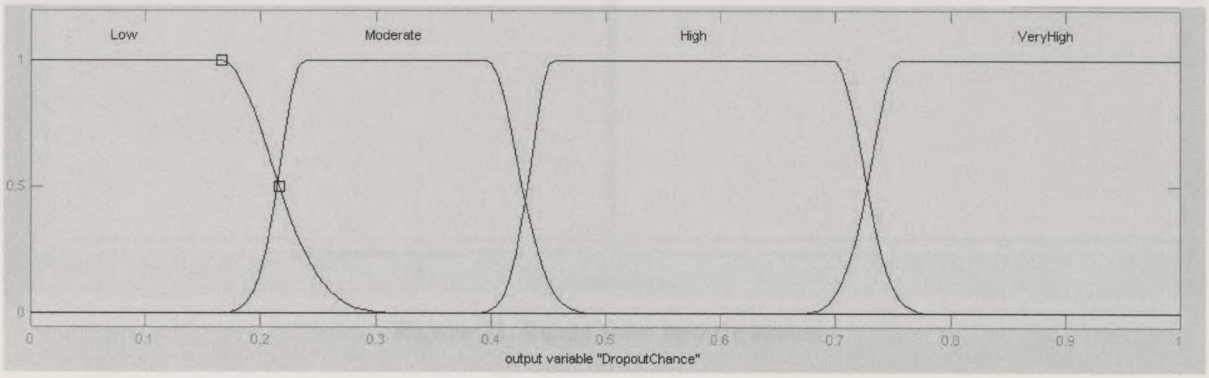

Figure 20. Fuzzy sets of output variable (DropoutChance)

Student age:

All records (retention and dropout) listed have ages between 16 and 40. There are 3 Gaussian fuzzy sets: Teen, Young and Older (Figure 21). The range is from 15 to 40.

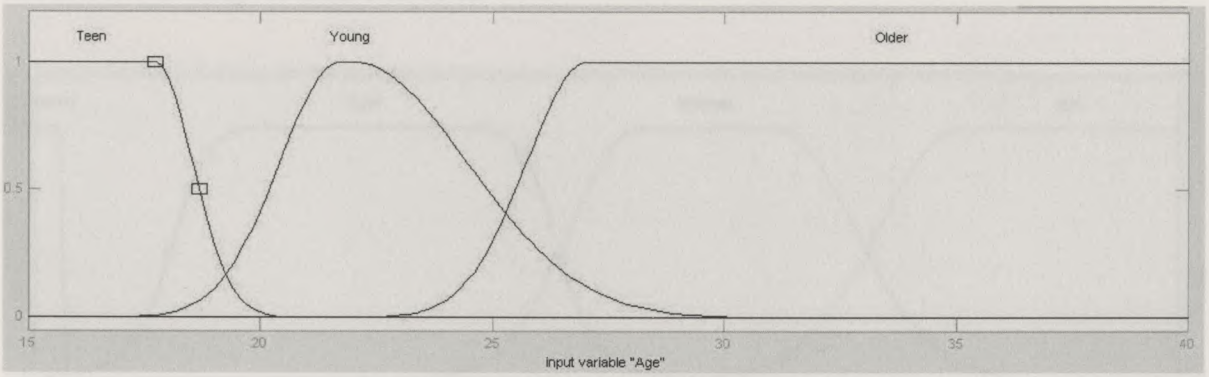

Figure 21. Fuzzy sets: Student age

Instate status:

The input variable Instate status has 2 Boolean sets: OutOfState and InState (Figure 22).

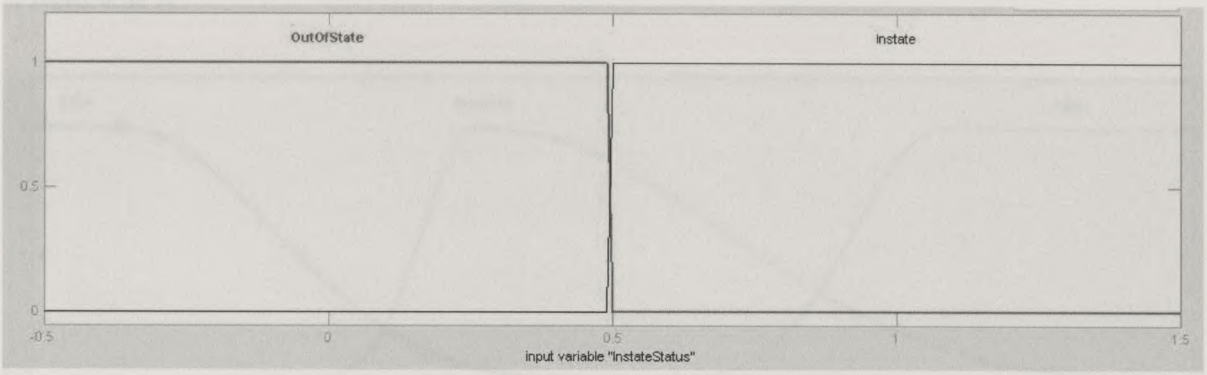

Figure 22. Fuzzy sets: Instate status

Student test score:

For student test score there are 3 Gaussian fuzzy sets: Low, Moderate and High (Figure 23). We also kept a set (Unknown) just in case we needed to filter out the records that have no student test score. The range is from 0 to 1600.

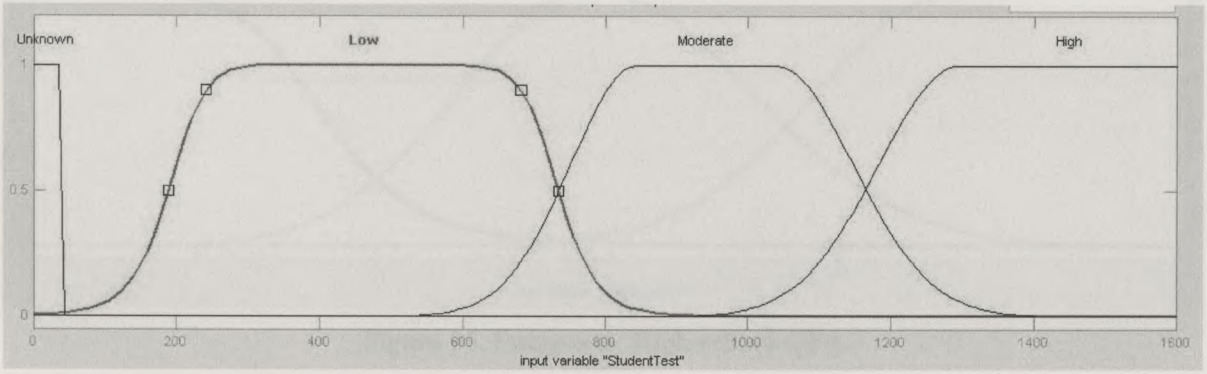

Figure 23. Fuzzy sets: Student test score

Course load:

For course load there are 3 Gaussian fuzzy sets: Low, Moderate and High (Figure 24). The range is from 0 to 7.

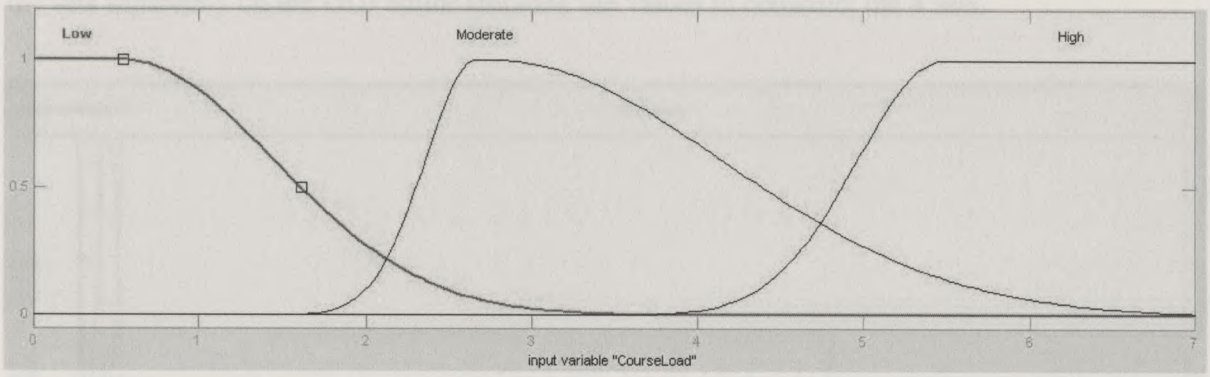

Figure 24. Fuzzy sets: Course load

High-school-GPA:

For High-school-GPA there are 3 Gaussian fuzzy sets: Low, Moderate and High (Figure 25). The range is from 2 to 4.

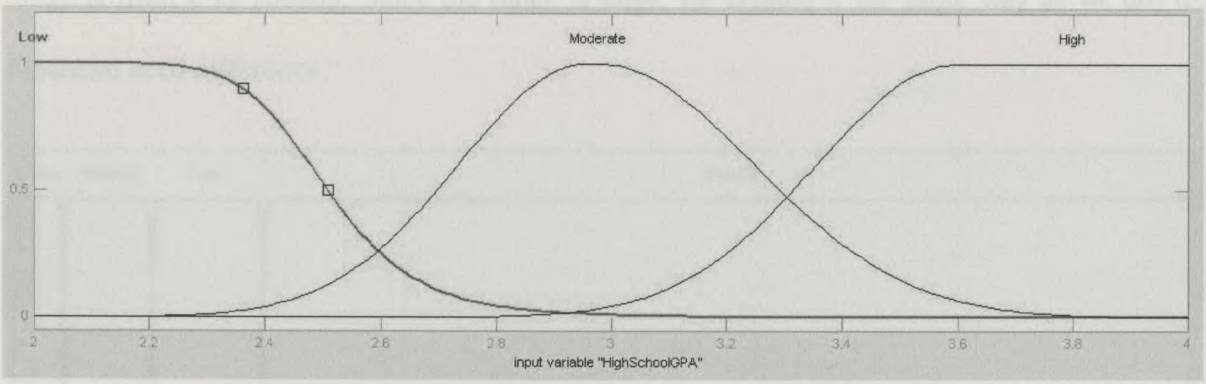

Figure 25. Fuzzy sets: High-school-GPA

Financial need difference:

For Financial need difference there are 4 Gaussian fuzzy sets: VeryWellOff, WellOff, Needy and VeryNeedy (Figure 26). The range is from -3000 to 40000. Since the range is so huge we created the sets separately on the GUI editor and took the values to construct the 4 sets.

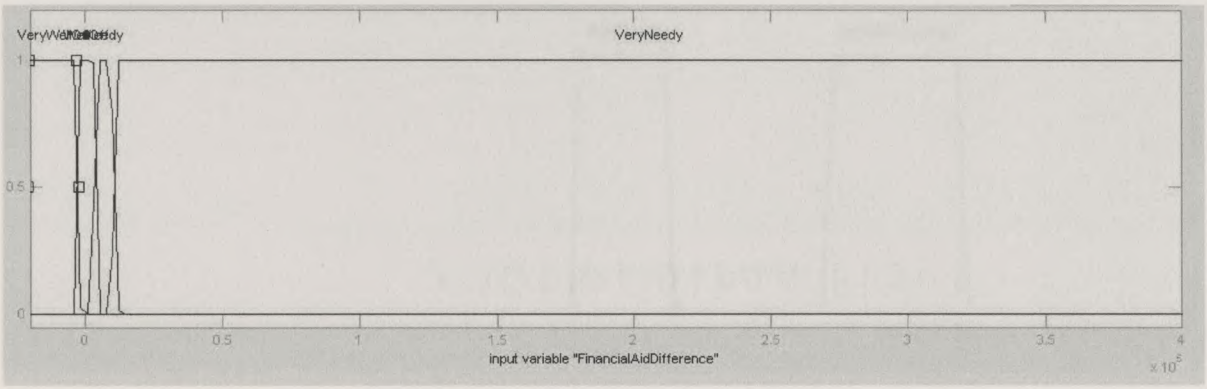

Figure 26. Fuzzy sets: Financial need difference

Estimated Family Contribution (EFC):

For EFC there are 4 Gaussian fuzzy sets: Low, Moderate, High and VeryHigh (Figure 27). The range is from 0 to 100000. Since the range is huge, we created it the same way as we did for Financial need difference.

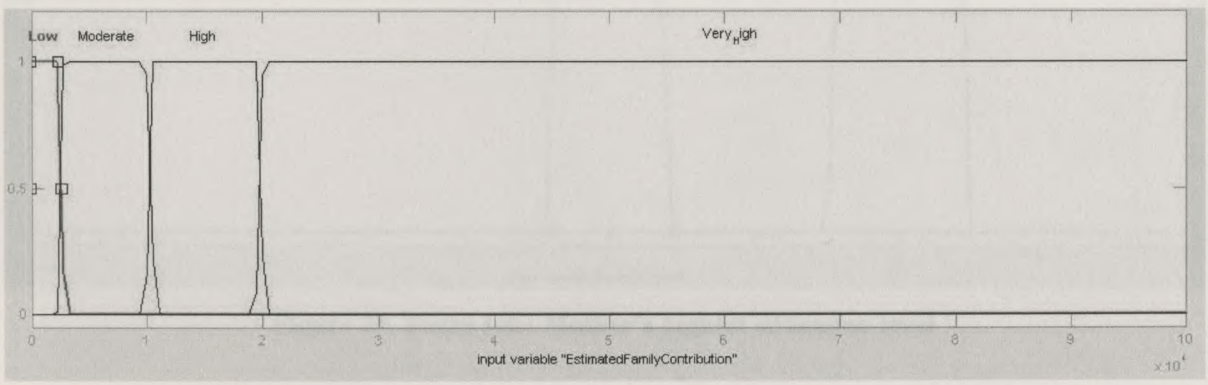

Figure 27. Fuzzy sets for input variable Estimated Family Contribution (EFC)

Father's highest education level:

We created 2 Boolean sets: HighSchool and MiddleSchoolJuniorHigh for this input variable (Figure 28). Since there are not a lot of records of students with parents having highest education level as College or beyond, we did not create a set for that.

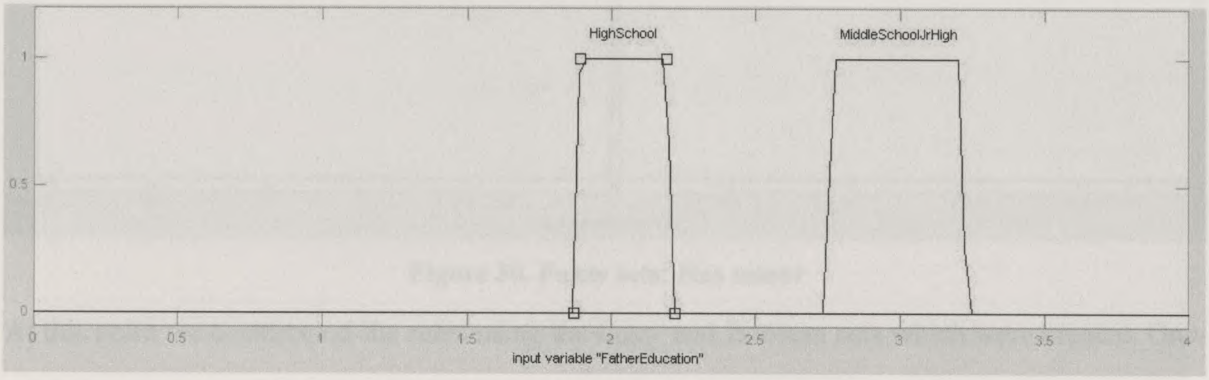

Figure 28. Fuzzy sets: Father's highest education level

Mother's highest education level:

For this input variable we followed the same methodology as we did for father's highest education level (Figure 29).

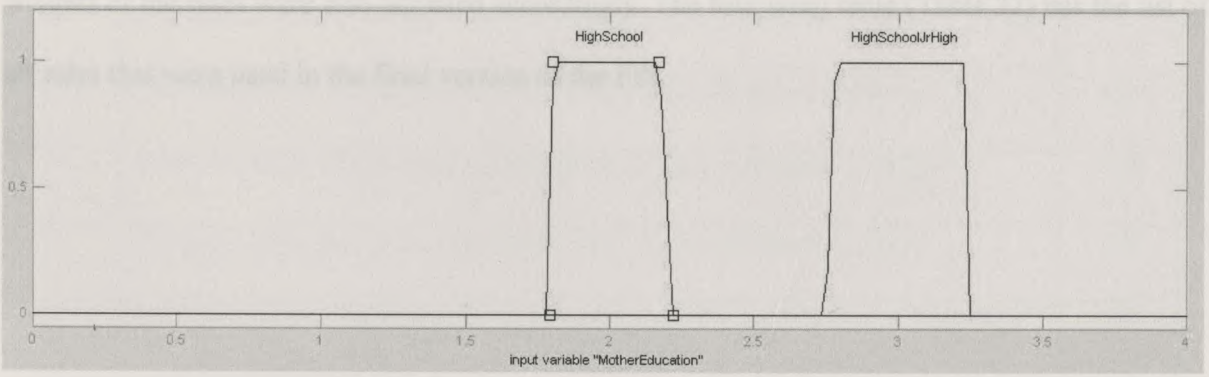

Figure 29. Fuzzy sets: Mother's highest education level

Has minor:

There are 2 Boolean sets: No and Yes for variable HasMinor as shown below in Figure 30.

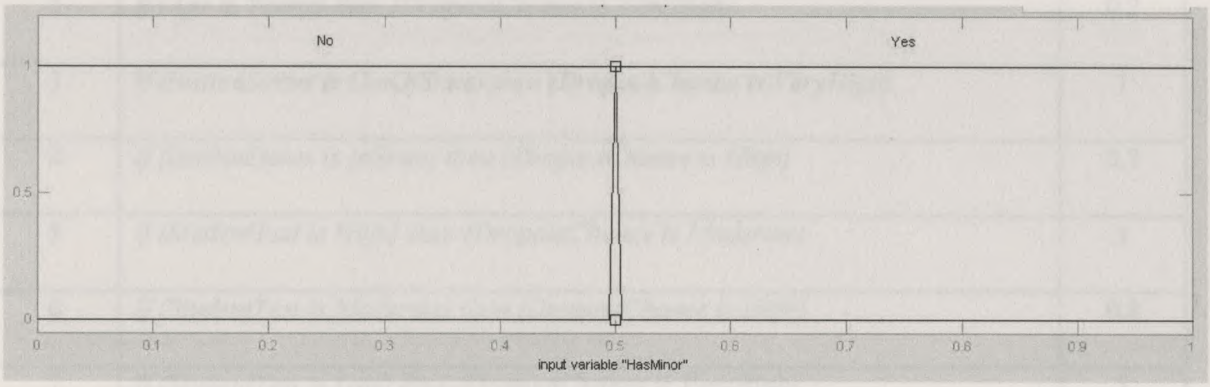

Figure 30. Fuzzy sets: Has minor

At this point we constructed the rules using the fuzzy and Boolean sets which were created. One-third of the retention and one-third of the dropout records were picked randomly for validating the system. The rest of the data was kept for testing. This data was converted into a format suitable for evaluating in Matlab. This was done using the utility program. Many rules that were included in the rule-base were left out after several trials as doing so yielded better results. The weights of the rules were also adjusted accordingly. The following table (Table 22) has the list of all rules that were used in the final version of the FIS.

Table 22. Rules used in the final version of FIS

Rule no.	Rule	Weight
1	*If (Age is Teen) then (DropoutChance is High)*	0.7
2	*If (Age is Young) then (DropoutChance is VeryHigh)*	0.7
3	*If (InstateStatus is OutOfState) then (DropoutChance is VeryHigh)*	1
4	*If (InstateStatus is InState) then (DropoutChance is High)*	0.7
5	*If (StudentTest is High) then (DropoutChance is Moderate)*	1
6	*If (StudentTest is Moderate) then (DropoutChance is High)*	0.8
7	*If (StudentTest is Low) then (DropoutChance is VeryHigh)*	1
8	*If (CourseLoad is Low) then (DropoutChance is VeryHigh)*	1
9	*If (CourseLoad is Moderate) then (DropoutChance is Moderate)*	0.8
10	*If (CourseLoad is High) then (DropoutChance is Low)*	0.8
11	*If (HighSchoolGPA is Low) then (DropoutChance is VeryHigh)*	1
12	*If (HighSchoolGPA is Moderate) then (DropoutChance is High)*	1
13	*If (HighSchoolGPA is High) then (DropoutChance is Low)*	1
14	*If (FinancialNeedDifference is VeryWellOff) then (DropoutChance is Low)*	0.8
15	*If (FinancialNeedDifference is WellOff) then (DropoutChance is High)*	0.8
16	*If (FinancialNeedDifference is Needy) then (DropoutChance is High)*	0.8
17	*If (FinancialNeedDifference is VeryNeedy) then (DropoutChance is VeryHigh)*	1
18	*If (EstimatedFamilyContribution is Low) then (DropoutChance is High)*	0.8
19	*If (EstimatedFamilyContribution is Moderate) then (DropoutChance is High)*	0.8
20	*If (EstimatedFamilyContribution is High) then (DropoutChance is Low)*	0.8
21	*If (EstimatedFamilyContribution is Very_High) then (DropoutChance is*	1

	Moderate)	
22	*If (FatherEducation is HighSchool) or (MotherEducation is HighSchool) then (DropoutChance is High)*	0.8
23	*If (FatherEducation is MiddleSchoolJrHigh) or (MotherEducation is HighSchoolJrHigh) then (DropoutChance is Moderate)*	0.6
24	*If (HasMinor is Yes) then (DropoutChance is Low)*	0.8

4. Experimental Results and Discussion

To evaluate the system's effectiveness we tested the FIS using the training data first. One-third of the available data was kept for training while the rest was kept for testing. The system was tweaked by changing the weights of the rules slightly to get better accuracy. The performance was then evaluated using the test data on the revised system. Finally a dataset created by merging the training and test data together was used for overall performance assessment. As discussed earlier, we initially set the threshold value of the output variable DropoutChance to 0.5. So any output value over this threshold was interpreted as a classification of dropout for a student.

4.1 Performance during Training

Table 23 shows the experimental results based on the training data. Although there were 630 actual dropouts among the total student population of 1969, it classified 999 students as likely dropouts. This amounted to a prediction accuracy of 39.1%.

Table 23. Results using training data

	Population	Percentage
Total	1969	
Number of actual dropouts	630 (32%)	
Number classified as dropouts	999	
Dropouts classified correctly	391	39.1%
Dropouts misclassified	608	
Percentage of actual dropouts classified correctly		62.1%
Percentage of actual dropouts misclassified		37.9%
Number of actual retentions	1339 (68%)	
Number classified as retentions	970	
Retentions classified correctly	731	
Retentions misclassified	239	
Percentage of actual retentions classified correctly		75.4%
Percentage of actual retentions misclassified		24.6%

In Figure 31 we can see that the correctly classified dropouts (the 39.1% mentioned above) represent 62.1% of the actual dropout population.

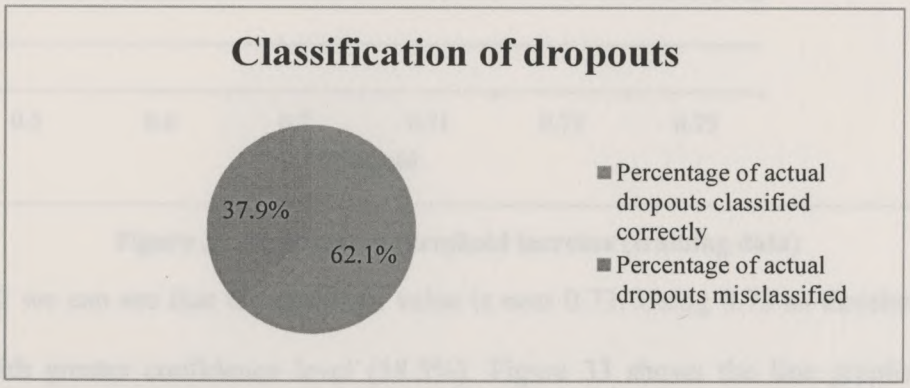

Figure 31. Classification of dropouts (using training data)

This rate of detecting dropouts is low. Students that are not actual dropouts are flagged as dropouts. To observe the effects of different threshold values for the output variable

DropoutChance on classification accuracy, the threshold value was varied in a range 0.5 to 0.73. Results of this experiment are shown in Table 24.

Table 24. Response to threshold increase (training data)

Threshold	Number classified as dropouts	Number of dropouts classified correctly	Confidence level
0.5	999	391	39.1%
0.6	509	237	46.6%
0.7	131	67	51.1%
0.71	114	58	50.9%
0.72	85	42	49.4%
0.73	65	38	58.5%

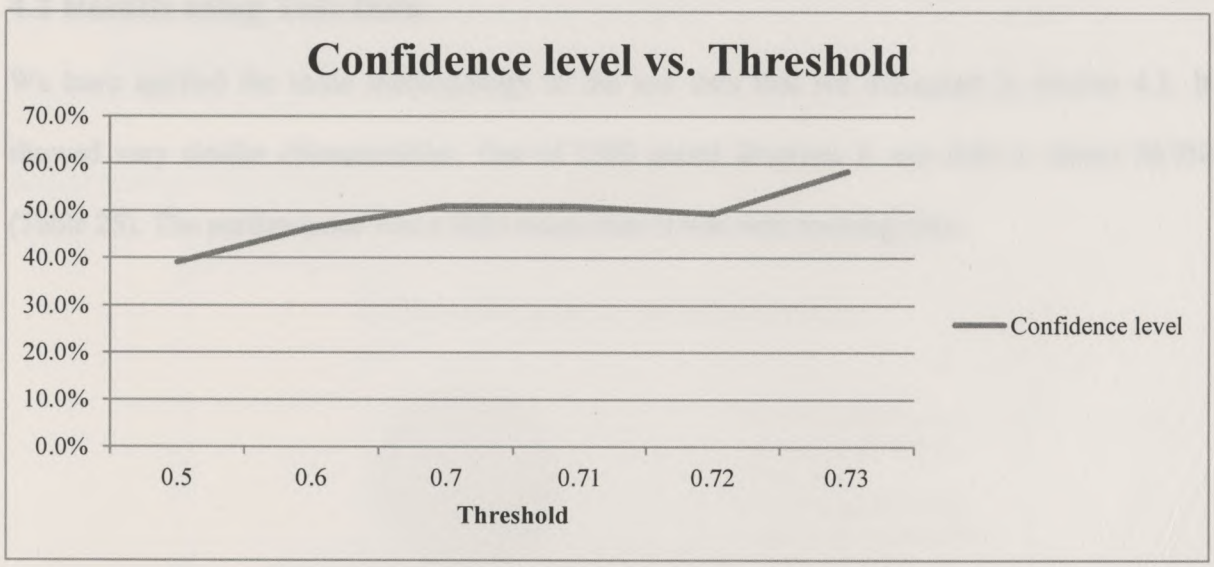

Figure 32. Response to threshold increase (training data)

In Figure 32 we can see that the optimum value is near 0.73. Using 0.73 as threshold yields 65 dropouts with greater confidence level (58.5%). Figure 33 shows the line graphs of number classified as dropouts and number of dropouts classified correctly against threshold.

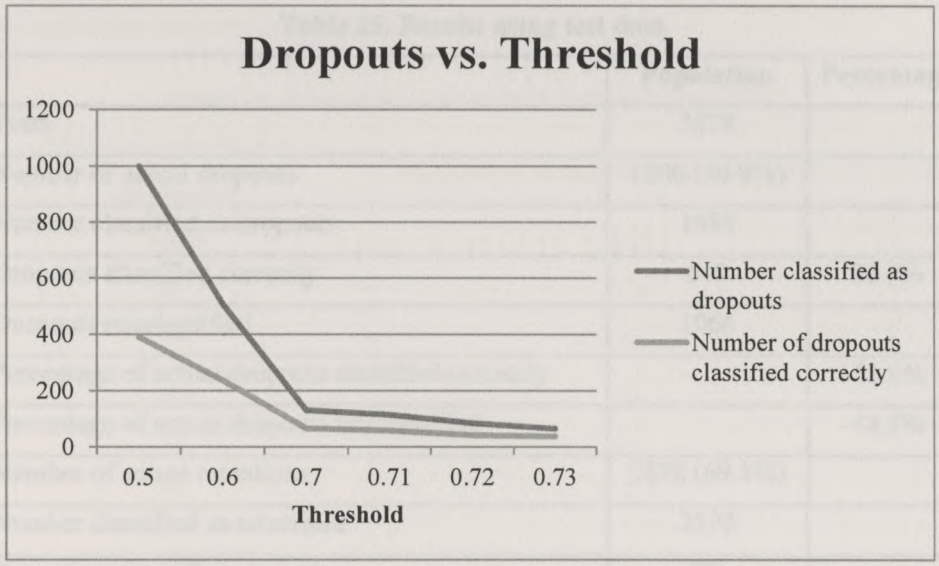

Figure 33. Line graphs: Dropouts vs. Threshold (training data)

4.2 Results using Test Data

We have applied the same methodology to the test data that we discussed in section 4.1. It showed very similar characteristics. Out of 1200 actual dropouts it was able to detect 36.7% (Table 25). The performance was a little worse than it was with training data.

Table 25. Results using test data

	Population	Percentage
Total	3878	
Number of actual dropouts	1200 (30.9%)	
Number classified as dropouts	1685	
Dropouts classified correctly	619	36.7%
Dropouts misclassified	1066	
Percentage of actual dropouts classified correctly		51.6%
Percentage of actual dropouts misclassified		48.4%
Number of actual retentions	2678 (69.1%)	
Number classified as retentions	2193	
Retentions classified correctly	581	
Retentions misclassified	1612	
Percentage of actual retentions classified correctly		26.5%
Percentage of actual retentions misclassified		73.5%

Actual dropouts classified correctly were 51.6% (Figure 34).

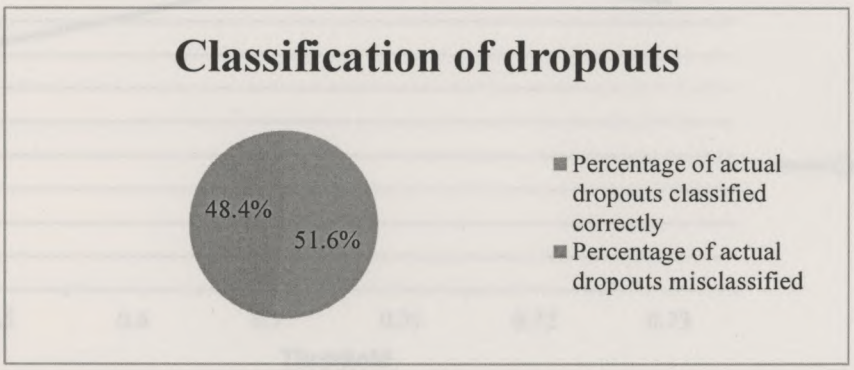

Figure 34. Classification of dropouts (using test data)

Response to threshold increase is shown in Table 26.

Table 26. Response to threshold increase (test data)

Threshold	Number classified as dropouts	Number of dropouts classified correctly	Confidence level
0.5	1685	619	36.7%
0.6	681	279	41.0%
0.7	125	58	46.4%
0.71	104	49	47.1%
0.72	80	37	46.3%
0.73	49	21	42.9%

In Figure 35 we can see that confidence level did not always increase as threshold increased. At 0.71 we have 85 correctly detected dropouts with a confidence level of 47.1%.

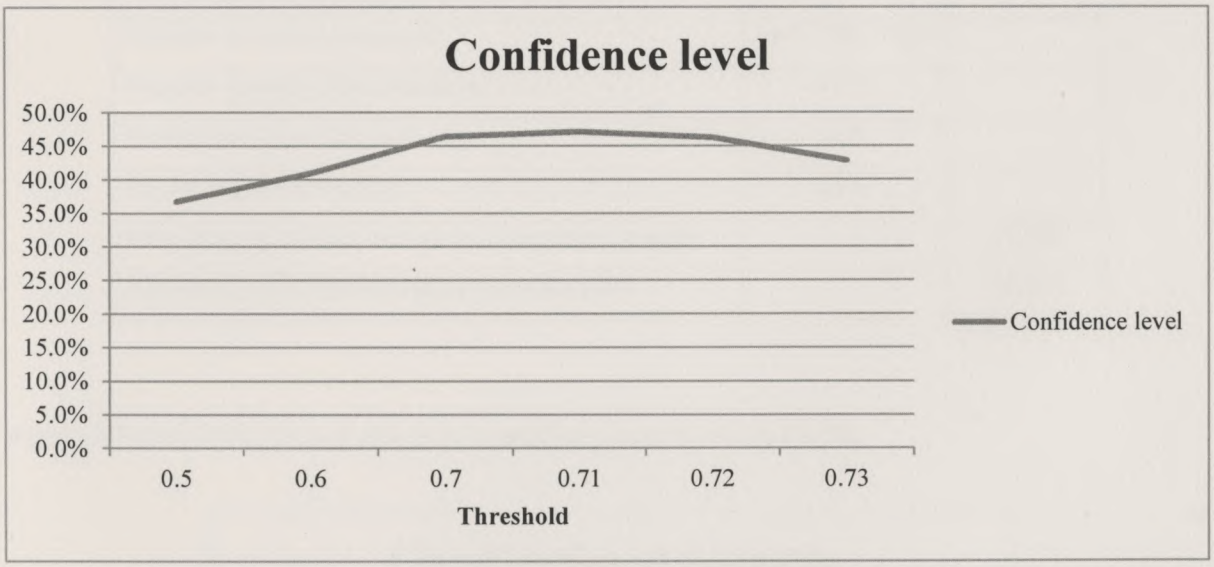

Figure 35. Response to threshold increase (test data)

In Figure 36 the line graphs of number classified as dropouts and number of dropouts classified correctly against threshold have been shown.

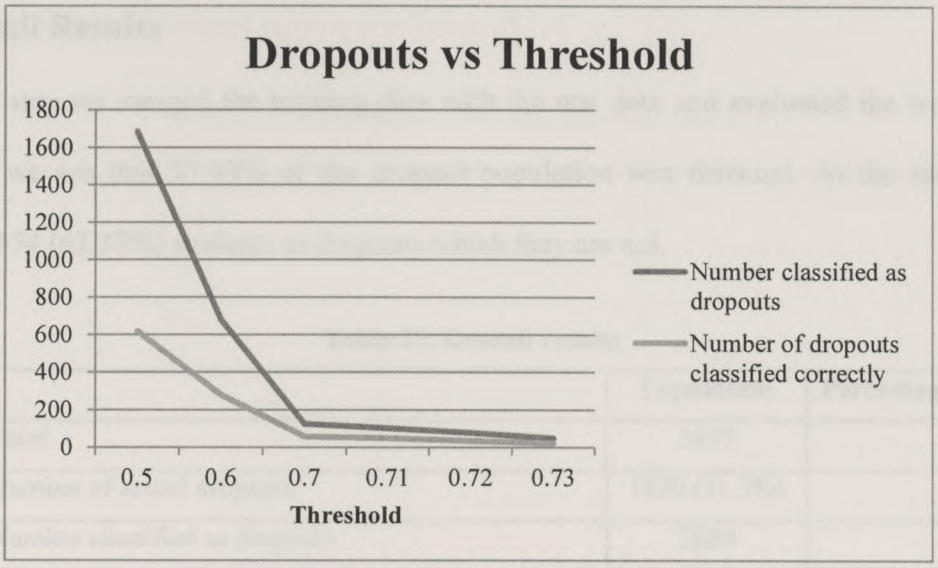

Figure 36. Line graphs: Dropouts vs. Threshold (test data)

4.3 Overall Results

In the last step we merged the training data with the test data and evaluated the results. From Table 27, we see that 37.63% of the dropout population was detected. At the same time it detected 1674 (62.37%) students as dropouts which they are not.

Table 27. Overall results

	Population	Percentage
Total	5847	
Number of actual dropouts	1830 (31.3%)	
Number classified as dropouts	2684	
Dropouts classified correctly	1010	
Dropouts misclassified	1674	
Percentage of actual dropouts classified correctly		55.2%
Percentage of actual dropouts misclassified		44.8%
Number of actual retentions	4017 (68.7%)	
Number classified as retentions	3163	
Retentions classified correctly	820	
Retentions misclassified	2343	
Percentage of actual retentions classified correctly		25.9%
Percentage of actual retentions misclassified		74.1%

Figure 37 shows that actual dropouts classified correctly were 55.2%.

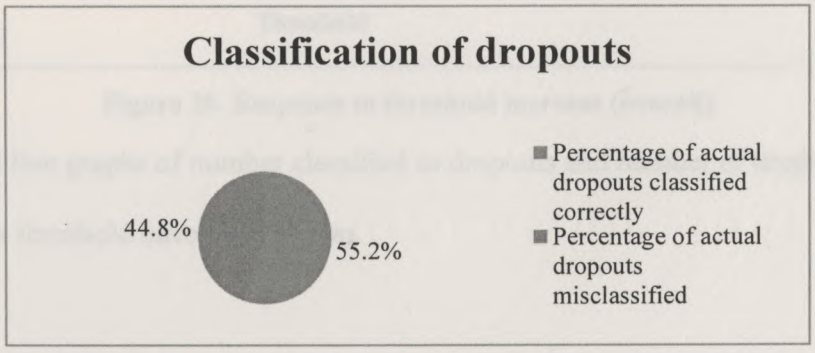

Figure 37. Classification of dropouts (overall)

Response to threshold increase is shown in Table 28.

Table 28. Response to threshold increase (overall)

Threshold	Number classified as dropouts	Number of dropouts classified correctly	Confidence level
0.5	2684	1010	37.6%
0.6	1190	516	43.4%
0.7	256	125	48.8%
0.71	218	107	49.1%
0.72	165	79	47.9%
0.73	114	59	51.8%

From Figure 38 we can see that at threshold value of 0.73 it gives us 59 dropouts that are flagged correctly. The confidence level was 51.8%.

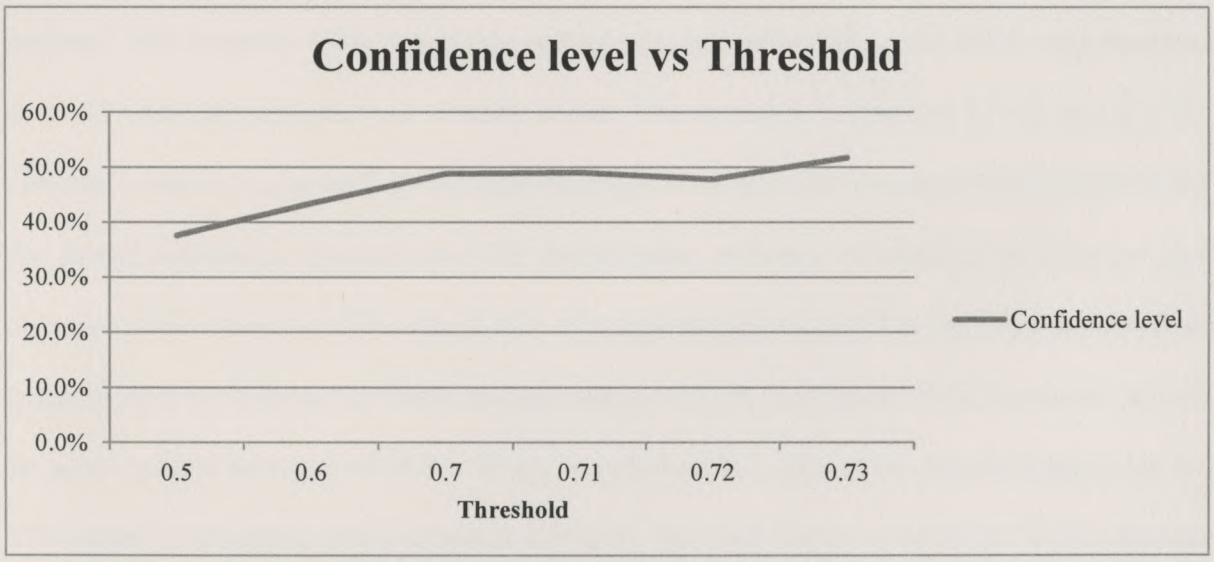

Figure 38. Response to threshold increase (overall)

In Figure 39 the line graphs of number classified as dropouts and number of dropouts classified correctly against threshold have been shown.

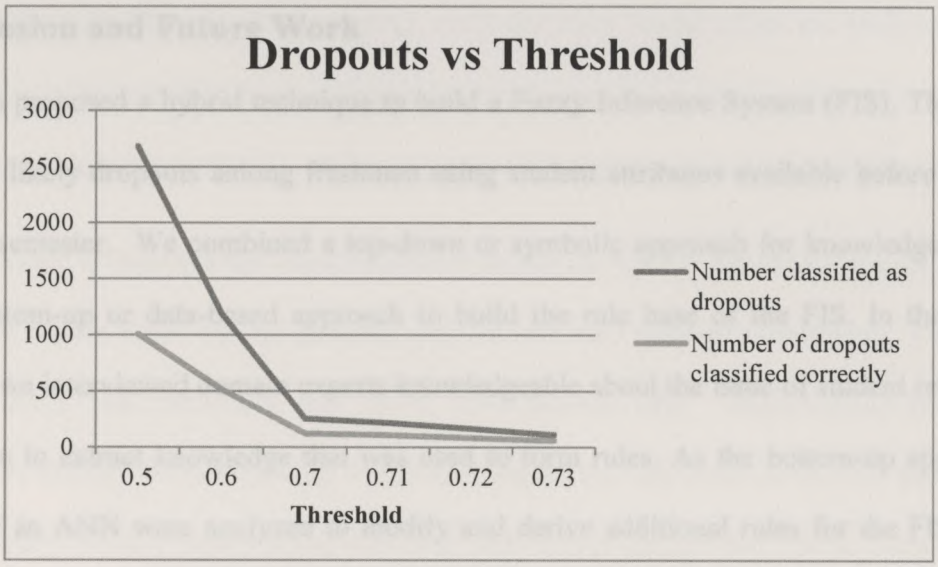

Figure 39. Line graphs: Dropouts vs. Threshold (overall)

5. Conclusion and Future Work

This thesis proposed a hybrid technique to build a Fuzzy Inference System (FIS). The goal was to classify likely dropouts among freshmen using student attributes available before the end of their first semester. We combined a top-down or symbolic approach for knowledge extraction with a bottom-up or data-based approach to build the rule base of the FIS. In the top-down approach, we interviewed domain experts knowledgeable about the issue of student retention and progression to extract knowledge that was used to form rules. As the bottom-up approach, the weights of an ANN were analyzed to modify and derive additional rules for the FIS. We also analyzed the student data to identify discriminating variables (student attributes), namely, variables whose values differed significantly for dropouts and returning students. . Variables that were found not be discriminating enough were left out of the FIS model. We also avoided using students' first semester GPA, a variable of high discriminating value, but which only becomes available after the semester has actually ended. This decision is justified by the fact that for remedial measures to be useful, they must be taken soon after the semester starts.Despite using the hybrid knowledge process, the FIS classification accuracy of students as dropouts was observed to be worse than that of the ANN. Although the percentage of actual dropouts classified correctly by the FIS for all available student data was 55.2% (see Table 26), this system can still be quite useful if deployed carefully. When we increased the value of the threshold applied to the FIS output, it classified fewer dropouts correctly, but with higher confidence levels. Another useful outcome of the hybrid approach used is that it produced rules relating student attributes to their dropout likelihood. A purely ANN based system, even though it classified the data with higher accuracy, behaves like a black box and does not provide any such insight,.

Relatively few significantly discriminating attributes were found when we analyzed the data. This leads us to believe that some attributes that might have played highly discriminating roles in determining a student's academic success were possibly not present in the available data set.

As future work to increase the classification accuracy of the FIS by improving its knowledge base, more expert views can be gathered and the rule base updated accordingly. The ANN we used for fuzzy rule extraction had about 75% success rate. Improvement in the ANN performance through further training would also impact the rule-base for the FIS and may yield better accuracy. Further tuning of the rules and fuzzy sets is another step that is likely to improve the FIS performance. Finally, there is also considerable potential for the development of a comprehensive decision support tool for detecting likely student dropouts by extending the utility program created during our experiment.

References

Andrews, R. Diederich, J., Tickle, A. (1995). *Survey and Critique of Techniques for Extracting Rules from Trained Artificial Neural Networks*. Knowledge-Based Systems, Vol. 8, pp. 373-389.

Chen, Z. (1999). *Computational Intelligence for Decision Support*. CRC Press.

Horstkotte, E & Togai. (1994). *InfraLogic, Inc.* Retrieved from http://www.austinlinks.com/Fuzzy/overview.html

Khan, S. (2011). *Notes on Fuzzy Systems for Artificial Intelligence course*, Columbus State University.

Khoo, L.P. & Zhai, L.Y. (2001). *Rclass*: A Prototype Rough-set and Genetic Algorithms Enhanced Multi-concept Classification System for Manufacturing Diagnosis*, CRC Press.

Klir, G. J., & Yuan, B. (1995). *Fuzzy sets and fuzzy logic* (pp. 487-499). New Jersey: Prentice Hall.

Kosba, E.; Dimitrova, V. & Boyle, R., (2003) *Using fuzzy techniques to model students in Web-based learning environment*, in Palade, V., Howlett, R.J. & Jain, L. (Eds.): *Knowledge-based Intelligent Information and Engineering Systems*, 7th International Conference, KES 2003, Oxford, United Kingdom, September 2003, Proceedings Part II, pp. 222–229.

Knapp, R. Benjamin. (2004). *Fuzzy Sets and Pattern Recognition*. Retrieved from http://www.cs.princeton.edu/courses/archive/fall07/cos436/HIDDEN/Knapp/fuzzy004.htm

Mamdani, E. H. (1974). *Application of fuzzy algorithms for control of simple dynamic plant*. Electrical Engineers, Proceedings of the Institution of, 121. Vol. 12, pp. 1585-1588.

MathWorks. (2012). *"What Is Mamdani-Type Fuzzy Inference?* - MATLAB & Simulink" Retrieved from http://www.mathworks.com/help/fuzzy/what-is-mamdani-type-fuzzy-inference.html

Mehra, N. (1973). *Retention and Withdrawal of University Students. (A Study of Academic Performance of a Freshman Class).*

Muslimi, B., Capretz, M. A., & Samarabandu, J. (2008). *An Efficient Technique for Extracting Fuzzy Rules from Neural Networks*. International Journal of Intelligent Technology, Vol. 1, pp. 3-4.

Nedic, Z.; Nedic, V. & Machotka, J. (2002). *Intelligent tutoring system for teaching 1st year engineering*, World Transactions on Engineering and Technology Education 2002, UICEE, Vol. 1, No. 2, pp. 241–244.

Pascarella, E. T., & Terenzini, P. T. (1979). *Interaction effects in Spady and Tinto's conceptual models of college attrition*. Sociology of Education, pp. 197-210.

Plagge, Mark. (2012). *Using Artificial Neural Networks to Predict First-Year Traditional Students Second Year Retention Rates.*

San Pedro, J. & Burstein, F. *Intelligent assistance, retrieval, reminder and advice for fuzzy multicriteria decision-making, in Palade*, V., Howlett, R.J. & Jain, L.C. (Eds.): KES 2003, pp. 37–44, Springer-Verlag, Berlin, Heidelberg.

Terenzini, Patrick T., Lorang, Wendell G., Pascarella, Ernest T. (1980). *Predicting freshman persistence and voluntary dropout decisions: A replication*. Research in Higher Education, Vol. 15, pp. 109-127.

Tinto, V. (1975). *"Dropout from higher education: A theoretical synthesis of recent research."* Review of Educational Research, Vol. 45, pp. 89-125.

Tsaganou et al. (2002). *Modeling student's comprehension of historical text using fuzzy case based reasoning*, Proceedings of the 6th European Workshop on Case-based Reasoning for Education and Training, Aberdeen, Scotland.

White, H. (1989). *Learning in artificial neural networks: A statistical perspective*. Neural computation, Vol. 1, pp. 425-464.

Xu, D.; Wang, H. & Su, K. (2002). *Intelligent student profiling with fuzzy models*, Proceeding of the 35th International Conference on System Sciences, pp. 1–8.

Yang, Y. (2005). *A conceptual framework for society-oriented decision support*, AI and Society, Vol. 19, pp.279–291, Springer-Verlag.

Yusof, Norazah., Ahmad, Nor Bahiah., Othman, Mohd. Shahizan., & Nyen, Yeap Chun. (2012). *A Concise Fuzzy Rule Base to Reason Student Performance Based on Rough-Fuzzy Approach.*

Yusof, N.; Mohd. Zin, N. A., Mohd. Yassin, N., & Samsuri P. (2009). *Evaluation of Student's Performance and Learning Efficiency based on ANFIS*, SocPar, pp. 460 – 465.

Zadeh, L. A. (1965). *Fuzzy sets*. Information and control, Vol. 8, pp. 338-353.

Appendix

Interview responses from domain experts:

Q1. Over the years, more female students dropped out of their study programs than their male counterparts. Do you think that female students are more likely to drop out? If yes, what can be the possible reasons?		
Expert 1 - Response	Expert 2 - Response	Expert 3 - Response
I agree, but usually they make comebacks. The possible reasons for dropouts are that they get married and need to look after children.	Not really. If the data suggest that then probably the reasons are having children, many of them are single parents. It's hard to find time for study.	No, I don't think so.
Q2. Do you think that compared with out-of-state students, in-state students are less likely to drop out?		
Expert 1 - Response	Expert 2 - Response	Expert 3 - Response
No, out of state students are more committed. Students from eastern parts are more likely to dropout. Students from Florida are very committed.	In-state students are more likely to drop out. Out-of-state students are more committed. The probable reasons behind out-of-state students are that many of them are from military.	Yes
Q3. Does financial aid play a positive role towards student retention? Are students who receive financial aid less likely to drop out?		
Expert 1 - Response	Expert 2 - Response	Expert 3 - Response

Yes. But it matters whether the financial aids are based on need or scholarship.	Strongly agree. Due to financial situation they are more interested about getting the monetary support. So they would want to retain the financial aid they receive.	Yes, I agree. Students are more committed when they have financial aid. Recently there were concerns about getting HOPE scholarships. Many students could not make it as the requirements were raised.
Q4. Do you agree with the notion that students failing in core subjects are more likely to drop out?		
Expert 1 - Response	Expert 2 - Response	Expert 3 - Response
Strongly agree. If students are not good in English, math and communication, it impacts on everything.	Strongly agree.	Agree.
Q5. Does parents' education level play a role towards student retention? Do you agree with the notion that students with college-educated parents are less likely to drop out?		
Expert 1 - Response	Expert 2 - Response	Expert 3 - Response
Probably there is more motivation. Not strongly agree though.	Yes. Strongly agree.	Yes
Q6. In the past, students with unmet financial need had higher dropout rates. Do you think unmet financial need can cause a student to drop out?		
Expert 1 - Response	Expert 2 - Response	Expert 3 - Response
Yes. Strongly agree.	Yes. Strongly agree.	Yes. Strongly agree.
Q7. Is high school GPA score a factor that positively correlates to retention?		
Expert 1 - Response	Expert 2 - Response	Expert 3 - Response

Yes. Better students stay in.	Yes. Agree.	I would say no. There are students with low GPA at high schools who did well afterwards.
Q8. Do you think that part-time students are more likely to drop out than full-time students?		
Expert 1 - Response	Expert 2 - Response	Expert 3 - Response
Yes, but they usually come back after taking breaks from their studies.	Yes. Many part-timers can't give commitment to study as they have jobs.	No. There are non-traditional students (aged and seniors) who are part-timers. Usually they don't drop out.
Q9. What in your opinion are the three most significant factors influencing student retention rates in your department?		
Expert 1 - Response	Expert 2 - Response	Expert 3 - Response
Ability to solve problems. Financial aid is important. More interaction with class-mates. ACM chapter these days is playing a positive role.	Desire to earn a degree. They understand the value of a degree but not education. Scheduling of classes. Many classes are offered online. Feeling safe at school. A few people don't want to face the real world challenges every day. They want to stay at school.	Financial aid. Serving communities.